PATCH2LOC: LEARNING TO LOCALIZE PATCHES FOR UNSUPERVISED BRAIN LESION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Detecting brain lesions as abnormalities observed in magnetic resonance imaging (MRI) is essential for diagnosis and treatment. In the search of abnormalities, such as tumors and malformations, radiologists may benefit from computer-aided diagnostics that use computer vision systems trained with machine learning to segment normal tissue from abnormal brain tissue. While supervised learning methods require annotated lesions, we propose a new unsupervised approach (Patch2Loc) that learns from normal patches taken from structural MRI. We train a neural network model to map a patch back to its spatial location within a slice of the brain volume. During inference, abnormal patches are detected by the anomaly score based on the error and variance of the location prediction. By applying the network in a convolutional manner, this generates a pixel-wise heatmap of anomalies providing finer-grained segmentation. We demonstrate the ability of our model to segment abnormal brain tissues by applying our approach to the detection of tumor tissues in MRI on T2-weighted images from BraTS2021 and MSLUB datasets and T1-weighted images from ATLAS and WMH datasets. We show that it outperforms the state-of-the art in unsupervised segmentation. *The code can be downloaded from the supplementary materials.*

1 INTRODUCTION

Detecting and localizing abnormal brain tissue in neuroimages is a critical diagnostic task, where early intervention can mitigate severe outcomes like those from incipient tumors or cortical malformations associated with epilepsy (Josephson et al., 2012). While Magnetic Resonance Imaging (MRI) is a powerful non-invasive modality for this task, modern supervised deep learning models are hindered by the scarcity of annotated data, particularly for rare and structurally diverse conditions (Busby et al., 2018; Hagens et al., 2019). This data bottleneck, combined with a shortage of neuroradiology experts (Merewitz & Sunshine, 2006), makes the development of robust unsupervised methods for anomaly detection an essential clinical and research goal.

Current state-of-the-art unsupervised approaches, including autoencoders (Zimmerer et al., 2018; Sato et al., 2019; Meissen et al., 2022b;a; Behrendt et al., 2022; Bercea et al., 2023a), adversarial autoencoders (Brock et al., 2018; Chen & Konukoglu, 2018; Baur et al., 2019), denoising autoencoders (Kascenas et al., 2022), transformers (Pinaya et al., 2022a) and Denoising Diffusion Probabilistic Models (DDPMs) (Wyatt et al., 2022; Pinaya et al., 2022b; Liang et al., 2023; Bercea et al., 2023a; Behrendt et al., 2024; 2025), operate on a principle of **global** reconstruction. These methods learn the typical structure of a normal brain and flag anomalies as regions where the model fails to reconstruct the input accurately. However, this reliance on global context might provide enough clues to reconstruct the abnormal regions (Wyatt et al., 2022; Bercea et al., 2023c; Kascenas et al., 2023). The best-performing models are DDPMs, but they suffer from a difficult trade-off called the **'noise paradox'** (Kascenas et al., 2023; Bercea et al., 2023b)—where enhancing anomaly signals can degrade the reconstruction of normal tissue, leading to false positives and limiting clinical reliability. These methods are thus constrained by a delicate balancing act of an inference-time hyperparameter (i.e., the amount of added noise) to be effective.

To overcome these limitations, we propose Patch2Loc, a novel framework that fundamentally shifts the paradigm from global reconstruction to local structural assessment. Instead of learning to reconstruct an entire image, Patch2Loc is trained on a simple yet powerful localizing task: predicting

the spatial coordinates of an isolated image patch from a normal brain. The intuition is that the model learns the distinct anatomical patterns characteristic of each location. When presented with a patch containing an anomaly, its structure deviates from the norm, causing the model to predict its location with high error and high uncertainty. This provides a direct and robust signal for anomaly detection based on localized structural failure. This is a key distinction from other location-based self-supervised learning (SSL) tasks such as Jigsaw (Noroozi & Favaro, 2016), patch ordering and context prediction (Doersch et al., 2015; Taleb et al., 2020), where the objective is to learn from the relationships between multiple patches (e.g., context or relative position). These tasks use the relationship between two or more patches within the global context. Their task losses are not localized to a single patch; thus, they were not designed for localized abnormality detection. In contrast, Patch2Loc can be applied convolutionally using overlapping patches to localize abnormalities by increases in the location prediction error and the model's uncertainty.

To summarize, Patch2Loc offers several key advantages:

1. **Spatially-Aware Local Feature Learning:** By learning to predict a patch's spatial origin, Patch2Loc encodes local anatomical patterns. This allows it to effectively distinguish abnormal patches based on deviations in location prediction, directly targeting the source of the anomaly.
2. **Uncertainty-Guided Anomaly Detection:** The variance in the model's predictions serves as a powerful uncertainty estimate. Combining the location prediction error with this uncertainty creates a more robust abnormality score that significantly improves detection performance.
3. **Minimal Inference Hyperparameter Dependence:** Unlike reconstruction-based methods that require careful tuning of noise levels during inference, Patch2Loc operates with a stable configuration, providing a pixel-level abnormality heatmap without complex post-training adjustments.

By focusing on local features and leveraging both prediction error and uncertainty, Patch2Loc offers an intuitive, interpretable, and effective solution for unsupervised anomaly detection with high potential for clinical applicability.

2 METHODOLOGY

The idea underlying Patch2Loc is that there is a strong relationship between patch content and location, due to the normal anatomical patterns in structural neuroimages, that is used by neuroradiologists when identifying structural abnormalities. With sufficient sampling across the population, this relationship can be modeled using machine learning. As each brain has slightly different sizes, the images are first registered using rigid transformation, such that the patches taken from the same coordinates contain similar anatomy. Figure 1 shows a summary schematic diagram for methodology behind Patch2Loc.

2.1 PROBLEM FORMULATION

Let $(Y_1, Y_2, A) \in [0, 100] \times [0, 100] \times [0, 100]$ denote the 3D location of a rectangular image patch $X \in \mathcal{X} \subset \mathbb{R}_{\geq 0}^{S_1 \times S_2}$ within the brain taken at the 2D location $Y = (Y_1, Y_2) \in [0, 100] \times [0, 100]$ from the slice located at $A \in [0, 100]$. The coordinates are percentages of the patch's location in absolute coordinates (L_1, L_2, L_3) relative to the brain volume's spatial extent along each axis (E_1, E_2, E_3): $Y_1 = 100 \cdot \frac{L_1}{E_1}$, $Y_2 = 100 \cdot \frac{L_2}{E_2}$, and $A = 100 \cdot \frac{L_3}{E_3}$. With registered brain scans, the spatial extents E_1, E_2, E_3 are constant for all scans. Each 2D patch is rectangular with an absolute size of $S_1 \times S_2$ chosen based on a fixed relative proportion $r = \frac{S_1}{E_1} = \frac{S_2}{E_2}$. The choice of r controls the patch's coverage.

The patch and location can be described as continuous random variables jointly distributed $(X, Y, A) \sim P$. Intuitively, $P(X|Y_1 = y_1, Y_2 = y_2, A = a)$ is the distribution of images at a particular location (y_1, y_2, a), but this is a high-dimensional distribution that is difficult to model. To capture the shared information between the image patch and its location, we model the conditional distribution $P(Y|X, A)$, which is two dimensional. For simplicity, we model this as a 2D

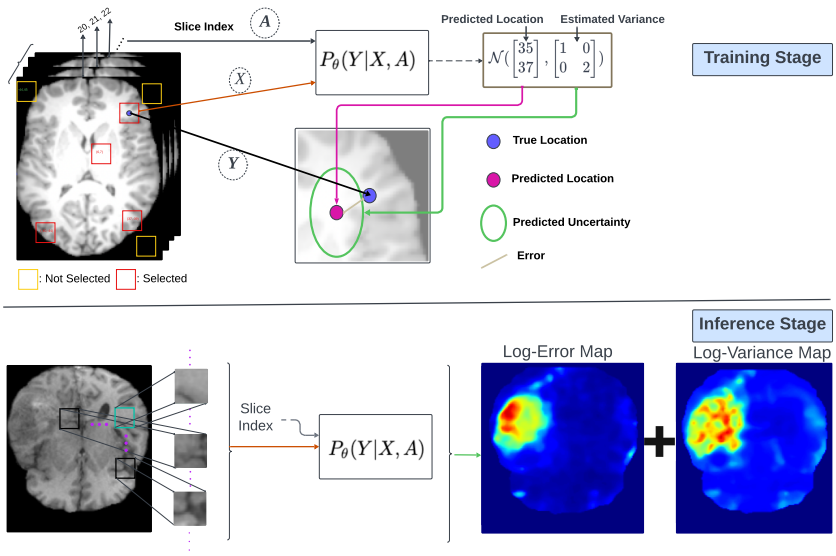


Figure 1: Schematic diagram for Patch2Loc. (Top) Training stage: A randomly selected patch from a normal MRI slice is paired with its two-dimensional location Y and slice index A. The relationship between the image patch, slice, and its location is modeled as a conditional distribution $P_\theta(Y|X,A)$ using a 2D Gaussian distribution defined by the mean and the variance of each coordinate that are functions of the patch and slice index. Note that patches with less than 20% of its content as brain tissues, as in the yellow patch, are rejected. (Bottom) Inference stage: Overlapping patches with a fixed stride are extracted as in convolution from an MRI slice and the model is applied to each patch. The squared norm of the error between the model's predicted mean and the patch's true location creates an error map. Likewise, the sum of the variances create a variance map. Together the sum of the logarithms of the errors and variances highlight anomalous areas.

Gaussian distribution,

$$P_\theta(Y|X,A) = \mathcal{N}(\mu^\theta(X,A), \Sigma^\theta(X,A)), \tag{1}$$

where $\mu^\theta(X,A)$ is the 2D mean and $\Sigma^\theta(X,A)$ is the covariance matrix, both are functions of the patch X and the slice location A, parameterized by θ. To further simplify the model, we consider a diagonal covariance matrix described by the variance of the two coordinates. The model is then $Y_i|X,A \sim \mathcal{N}(\mu_i^\theta(X,A), \Sigma_{ii}^\theta(X,A)), \quad i \in \{1,2\}$.

The functions predicting the conditional mean and variances μ^θ and Σ^θ of location given the patch are modeled using a neural network with parameters $\theta = (\theta_0, \theta_m, \theta_v)$, where θ_0 denotes the parameters of the shared patch encoder $\phi : \mathcal{X} \to \mathbb{R}^d$, θ_m denotes the parameters of the mean prediction head $\tilde{\mu}$, such that $\mu^\theta(X,A) = \tilde{\mu}(\phi(X),A) \in \mathbb{R}^2$, and θ_v denotes the parameters of log-variance head $\tilde{\varsigma}$, such that $\varsigma^\theta(X,A) = \tilde{\varsigma}(\phi(X),A) \in \mathbb{R}^2$ and $\Sigma_{ii}^\theta(X,A) = \exp(\varsigma_i^\theta(X,A)), \quad i \in \{1,2\}$.

2.2 Loss Function

The model parameters could be optimized to minimize the negative-log likelihood (NLL)

$$\theta^* \in \arg\min_\theta \mathbb{E}_{(X,Y,A)\sim P}[-\log P_\theta(Y|X,A)]. \tag{2}$$

However, using NLL for estimating a model of both the mean and the variance is problematic as the component of the loss with respect to the mean estimates are scaled by the variance estimates,

$$-\log P_\theta(Y|X,A) \propto \sum_{i=1}^{2} \frac{|Y_i - \mu_i^\theta(X,A)|^2}{\exp(\varsigma_i^\theta(X,A))} + \varsigma_i^\theta(X,A), \tag{3}$$

which causes convergence to local minima with poor performance (Nix & Weigend, 1994; Seitzer et al., 2022). Hence, we adapt the β-NLL loss proposed in the work by Seitzer et al. (2022) that

scales each dimension of the loss by the variance estimate taken to the βth power, $\beta \in (0, 1]$,

$$\sum_{i=1}^{2} \lfloor \exp(\varsigma_i^\theta(X, A)) \rfloor^\beta \left(\frac{|Y_i - \mu_i^\theta(X, A)|^2}{\exp(\varsigma_i^\theta(X, A))} + \varsigma_i^\theta(X, A) \right), \tag{4}$$

where $\lfloor \cdot \rfloor$ indicates the stop-gradient operation such that gradients are not taken with respect to this scaling. This scaling mitigates the effect of the variance estimate on the gradient of the loss with respect to the mean. Notably, when $\beta = 1$ then μ_i^θ is updated based on mean-squared error. Thus, we choose the suggested value of $\beta = 0.5$ (Seitzer et al., 2022).

2.3 Abnormality Score

To detect abnormalities with Patch2Loc, we calculate an abnormality score from the location prediction error and the sum of the variance estimates, after a log transformation. The log-error is computed as $\log(\|Y - \mu^\theta(X, A)\|_2^2 + \varepsilon)$, where $\varepsilon = 0.5$ such that it does not effect large errors. The log-variance is computed as $\frac{1}{2}\log(\det(\Sigma^\theta(X, A))) = \frac{1}{2}\sum_{i=1}^{2}\log(\Sigma_{ii}^\theta(X, A))) = \frac{1}{2}\sum_{i=1}^{2}\varsigma_i^\theta(X, A)$. The log-variance captures the model's uncertainty since adding $1 + \log(2\pi)$ to the log-variance is the entropy of $\mathcal{N}(\mu^\theta(X, A), \Sigma^\theta(X, A))$. The abnormality score is

$$\text{Score}(X, Y, A) = \underbrace{\log(\|Y - \mu^\theta(X, A)\|_2^2 + \varepsilon)}_{\text{Log(Error}^2\text{)}} + \underbrace{\frac{1}{2}\varsigma_1^\theta(X, A) + \frac{1}{2}\varsigma_2^\theta(X, A)}_{\text{Log(Variance)}}. \tag{5}$$

It assigns the highest values to data points that are both far from the predicted mean and are associated with a large predicted variance. Essentially, when the model expects being wrong and is. Normal patches where the model is certain and have low error predictions have the lowest anomaly scores.

2.4 Model Details

While the Patch2Loc task can be applied to different anatomical orientations for the slices, we consider axial slices, such that A is indicates the location of the slice along the vertical dimension (bottom to top of the brain). We select $r = 12.5\%$ for the proportion of the axial slice's width and length. Ideally, the smaller the patch, the more precise the resulting heatmap would be. However, in development we found that smaller patch sizes deteriorates Patch2Loc's performance since the patches are more ambiguous. When the patch size is too small, the inherent structural symmetry of the brain causes patches from different locations to appear similar, making it challenging for Patch2Loc to distinguish between them effectively. Contrastingly, larger patches may fail because the model can use the context surrounding an abnormality to accurately predict a location.

For the encoding model $\phi : \mathbb{R}_{\geq 0}^{S_1 \times S_2} \to \mathbb{R}^d$, which serves as a common backbone, we use ResNet-18 (He et al., 2016), producing a shared latent feature vector with dimension $d = 512$. Since the patch size are small (i.e., $S_1 \times S_2 = 24 \times 24$), we replace the ResNet-18 first convolution layer with a kernel of size 3 instead of 7. The latent vector from the backbone $\phi(X) \in \mathbb{R}^d$ is added to a sinusoidal positional encoding of the slice coordinate A (Vaswani, 2017). This representation is input to two separate branches for the mean $\tilde{\mu} : \mathbb{R}^d \to \mathbb{R}^2$ and the log-variance $\tilde{\varsigma} : \mathbb{R}^d \to \mathbb{R}^2$. Each branch consists of four fully-connected layers with output dimensions $512, 128, 64, 32$, with a batch normalization layer and then a rectified linear unit (ReLU) activation function after each fully-connected layer, followed by a final linear layer to the two-dimensional outputs.

2.5 Convolutional Approach for Pixel-level Detection

To perform pixel-level abnormality detection, we employ Patch2Loc in a convolutional manner. Specifically, we feed overlapping patches with a specified stride into the Patch2Loc model, padding patches with zeros when necessary. Subsequently, we utilize the Patch2Loc outputs, predicted mean and variance, along with the corresponding true coordinates to construct the error, variance, and abnormality score map. This is depicted in the bottom section of Figure 1. When the stride is 1, the scoring map has the same resolution as the input MRI image. If a larger stride is used, the abnormality score heatmap will have a lower resolution, but it can be upsampled (e.g., via interpolation) to match the input dimensions. Throughout all experiments, we use a stride of 1.

3 Related Work

While not used for unsupervised abnormality segmentation, one prior work by Taleb et al. (2020) uses self-supervised learning with a task that resembles Patch2Loc's location prediction task. The task is a 3D version of a previous context-dependent self-supervised task proposed by Doersch et al. (2015). Specifically, the task consists of predicting the discrete relative location of a 3D patch among the possible locations in a $3 \times 3 \times 3$ grid surrounding a center patch that is provided as input. The performance for unsupervised abnormality segmentation was not benchmarked as the task served only as a pretext task to pretrain a backbone network before subsequent supervised segmentation (Taleb et al., 2020). Patch2Loc is distinguished in that its training directly informs the abnormality score. Additionally, Patch2Loc predicts the continuous location of a patch without the need of a center patch as context, using only the slice index as sufficient context for registered brain images.

Unsupervised abnormality segmentation in neuroimaging has benefited from improving machine learning models ranging from autoencoders (AE), variational autoencoders (VAE) (Baur et al., 2021), generative adversarial neural networks (GANs) or discriminators as in GANs (Goodfellow et al., 2014), transformers, and diffusion models. Nonetheless, a notable baseline is based on a simple threshold on the image histogram (Meissen et al., 2022a), which can exceed the performance of baseline methods for certain modalities. Here we describe works relevant to Patch2Loc or those we compare against, a more comprehensive discussion of the related work can be found in appendix A.1.

Many approaches leverage AE or VAE in novel ways. Rather than relying on squared error for training, another work by Meissen et al. (2022b) uses an AE architecture that reconstructs features obtained from a pretrained encoder using the Structural Similarity Index Measure (SSIM) as the loss function. SVAE by Behrendt et al. (2022) uses a VAE with transformers to capture the inter-slice dependencies and showed it can improve the results compared to 2D vanilla VAE. The RA method by Bercea et al. (2023c) uses an VAE with a cyclic loss and use the reconstruction error as abnormality score. Baur et al. (2021) noted that AE and VAE methods suffer from blurry reconstructed images, which hinders their performance. Incorporating a discriminator as in a GAN can improve the reconstruction quality.

One prior work by Van Hespen et al. (2021) used a patch-based auto-encoder with a cycle consistency term and a discriminator to distinguish between the original image and its reconstruction, testing it on specific abnormal tissues. We identified it as the only approach leveraging local features for unsupervised abnormality segmentation in brain MRI. However, the method was applied to a dataset of infarcts (dead brain tissue caused by loss of blood flow) and was not compared for abnormal tissue detection.

Another approach is to iteratively restore an image to better match the normal data distribution. PHANEs (Bercea et al., 2023d) uses a model to restore part of an MRI slice flagged by the RA method (Bercea et al., 2023c) to mitigate false positives within the flagged region. This is along the lines of denoising autoencoder (DAE) (Kascenas et al., 2022), which learns to remove correlated noise added to the input images during training. DAE outperformed GAN and VAE approaches, achieving higher Dice score and average precision.

Denoising diffusion probabilistic models (DDPM) (Ho et al., 2020) build on DAEs. One of the first works to use DDPM (Wyatt et al., 2022) for unsupervised abnormality segmentation also proposed learning to denoise simplex noise instead of Gaussian noise to enhance performance. While this empirically works, how this matches the fundamental assumptions of diffusion processes is not clear. Specifically, simplex noise is procedurally generated, and is not described by a random process. In contrast, a diffusion model's forward process is described by a Markov chain, with a Markov transition kernel (Sohl-Dickstein et al., 2015), typically described by adding Gaussian noise. Nonetheless, the simplex noise version of DDPM formulation essentially defines a training regime for denoising across different signal-noise levels, where longer times correspond to higher noise regimes. During inference, denoising is performed from the initially high noise regime, then new simplex noise is applied at decreasing noise levels, and the denoising process continues. The patched diffusion model (pDDPM) by Behrendt et al. (2024) performs add noise only to a patch of the whole slice. Different versions of each slice with patches at different locations are used to identify abnormalities within the slice. Behrendt et al. (2025) used conditional DDPM (cDDPM) to denoise a slice given the latent embedding from a masked auto-encoder (MAE) (He et al., 2022) pretrained on normal MRI slices, which is further fine-tuned during training.

4 DATA

We preprocess the neuroimages by sequentially applying the following processes: skull-stripping, registering each to the SRI atlas (Rohlfing et al., 2009), resampling to a voxel dimension of 1 mm^3 in the atlas space, applying N4 bias-correction, applying the histogram standardization method proposed by Nyul et al. (2000), and dividing each MRI image by its 98th percentile. In the penultimate step, the histogram standardization method in (Nyul et al., 2000) uses statistics (e.g., quantiles and the second mode) obtained from a dataset. We use the training dataset to estimate these statistics, and then we standardize every image from all datasets using the same statistics.

For the training data, we use the IXI dataset (Biomedical Image Analysis Group, Imperial College London, 2015), which contains MRI scans in both T1- and T2-weighted modalities for 560 subjects. Following the procedure outlined in Behrendt et al. (2024; 2025), a total of 161 samples are set aside for testing, while the remaining data is divided into five sets for cross-validation. Each set consists of 358 training samples and 44 validation samples.

For evaluation, we employ three different datasets: BraTS21 (Menze et al., 2014; Baid et al., 2021) with 1152 subjects; multiple sclerosis patients with lesion segmentation based (MSLUB) (Lesjak et al., 2018) with 30 subjects, white matter hyperintensity (WMH) (Kuijf et al., 2019) with 60 subjects, and Anatomical Tracings of Lesions After Stroke v2.0 (ATLAS) (Liew et al., 2022) with 955 subjects. The first two use T2-weighted and the last two use T1-weighted modality. These datasets represent different types of abnormal tissues. BraTS21 focuses on brain tumors with varying sizes and convex structures. MSLUB and WMH contain lesions from multiple sclerosis and white matter hyperintensities, respectively, which are relatively smaller and more scattered than tumors. ATLAS has small convex shaped abnormal tissues. Sample slices from these datasets, along with their ground truth abnormal tissue segmentation, are shown in Figure 4.

4.1 PATCH2LOC TRAINING DETAILS

For each training batch, we randomly select one slice from each of the 358 training subjects. From the combined pool of 358 slices, we uniformly sample 8096 patches. Patches are discarded if more than 80% of their pixels are background (i.e., $<$20% brain tissue). This is depicted in the top section of the schematic diagram shown in Figure 1, where the yellow outline patch in the top left corner are rejected due to their substantial empty content. This process defines a single batch, and Patch2Loc is trained for 15,000 such batches. Adam (Kingma & Ba, 2014) is used as an optimizer with a learning rate 10^{-2} and other hyper-parameters are left as defaults. We use a single NVIDIA V100 GPU to train the model.

5 RESULTS

To better understand Patch2Loc's operation, we investigate the distributions of Patch2Loc's location prediction errors and predicted variances, underlying the abnormality score, across normal and abnormal patches. Then we qualitatively illustrate Patch2Loc's abnormality heatmap on representative examples from each dataset. Finally, we compare Patch2Loc with benchmark and state-of-the-art (SOTA) methods (Meissen et al., 2022b;a; Behrendt et al., 2022; Bercea et al., 2023a;d; Wyatt et al., 2022; Behrendt et al., 2024; 2025) that reflect different methodologies such AE, VAE, GANs, and diffusion models that have published results on these datasets.

5.1 ABNORMALITY SCORE ANALYSIS

We illustrate the operation of Patch2Loc by examining the predicted distribution (mean and variance) of patches extracted from the same spatial location across subjects. For patch-level analysis, we obtain a set of non-overlapping patches from the abnormal datasets across all slices and individuals. We categorize a patch as either normal if it contains less than 10% of abnormal tissues or abnormal if it comprises over 90% of abnormal tissues. Patches falling within the 10% to 90% abnormal tissue percentage range are separately analyzed. We compare Patch2Loc's predictions for normal and abnormal patches drawn from the BraTS dataset in Figure 2. For normal patches, we observe low uncertainty (i.e., smaller ellipses) and low error, as the predictions are tightly clustered around the true location. Conversely, for abnormal patches, the model exhibits higher uncertainty

(i.e., larger ellipses) and greater prediction error, with the predicted means deviating significantly from the true location.

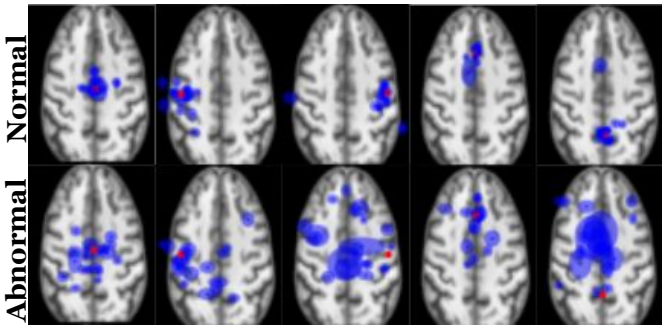


Figure 2: Visualization of Patch2Loc's output (blue ellipses) for patches captured from the same location (red dot) across different subjects in the BraTS dataset, overlaid on a representative T1-weighted registered slice (without abnormalities). (Top row) Predictions of normal patches. (Bottom row) Corresponding predictions of abnormal patches. The predicted Gaussian distribution is visualized as an ellipse, where the center represents the predicted mean, and the major and minor axes correspond to two standard deviations.

Figure 3 shows kernel density estimates (KDEs) of Patch2Loc's location prediction log-errors and predicted log-variance on normal and abnormal patches. (Figure 5 in the Appendix shows the results for all datasets). There is a clear separation between the normal and abnormal patches in the space of

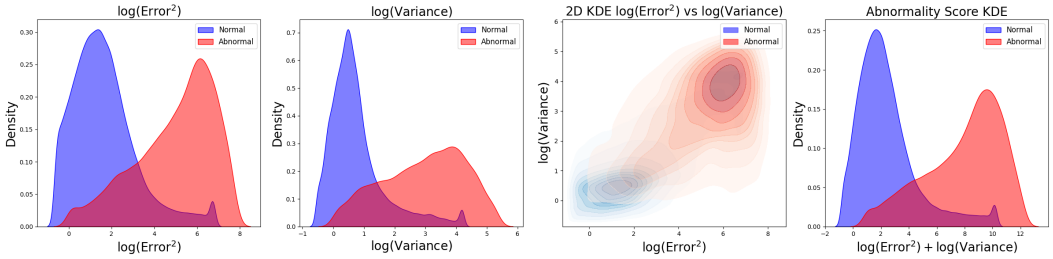


Figure 3: From left to right: 1D KDE for log-error, and log-variance, 2D KDE for log-error and log-variance, and 1D KDE for the abnormality score for normal patches (blue) and abnormal patches (red) for the BraTS dataset.

log-error and log-variance although there is some overlap. This overlap occurs because Patch2Loc accurately predicts the location of abnormal patches with low variance. Such predictions are more likely when the abnormal patches are located at the edges of the brain, where Patch2Loc can utilize the surrounding empty space to make precise location predictions. This is mitigated by having overlapped patches as explained in Section 2.5.

To investigate Patch2Loc on the partially abnormal patches (i.e., the patches that have abnormal content between 10% and 90%), we calculate the Spearman correlation between the abnormal content and log-error^2, log-variance, and the abnormal score (i.e., their sum) for each dataset as shown in Table 1. In all datasets, the abnormality score of the sum better correlates with the level of proportion of abnormal tissue in a patch compared to either log-error or log-variance alone, highlighting their complementary information.

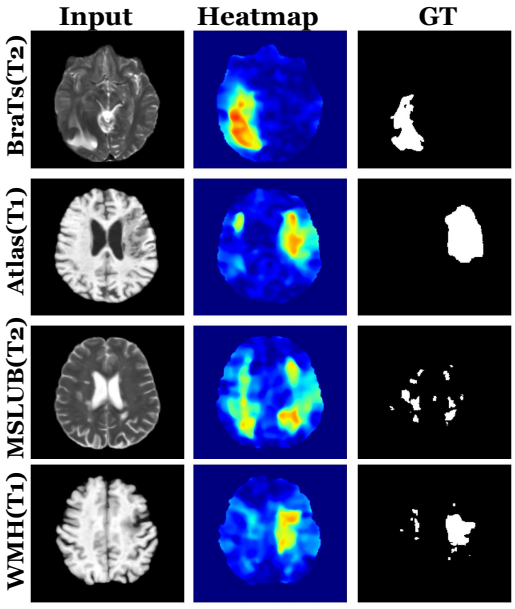
5.2 Qualitative Analysis

In Figure 4a, we present visualizations of MRI slices from the BraTS, ATLAS, MSLUB and WMH datasets, along with the corresponding heatmaps and ground truth annotations for abnormal tissues. While only one representative slice from each dataset are shown, they reflect a consistent pattern

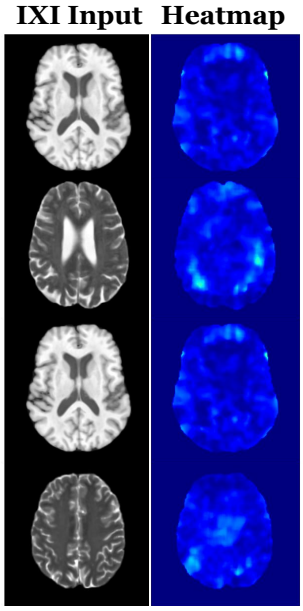
Table 1: Spearman correlation (%) between a patch's portion of abnormality and the log-error, log-variance, or abnormality score.

Metric	BraTS	MSLUB	ATLAS	WMH
Log-error	37%	15%	20%	12%
Log-variance	34%	14%	18%	11%
Score	40%	17%	23%	13%

observed across both datasets. Additional visualizations are provided in the Appendix (Figure 7 for BraTS, Figure 8 for MSLUB, Figure 9 for ATLAS, and Figure 10 for WMH). We also provide visualizations for slices from IXI test set (i.e. normal subjects) where it shows low abnormal scores in Figure 4b with more slices in Figure 6 in the Appendix. For fair comparison, the colormap is fixed to [0, 12], as the histograms show scores are typically concentrated within this range, even though lower/higher values may occur.



(a) Anomaly Detection on Pathological Cases



(b) Model Response on Healthy Controls

Figure 4: Visualization of our model's anomaly detection performance. **(a)** On pathological slices from the BraTS, ATLAS, MSLUB, and WMH datasets, the model's heatmap successfully localizes abnormalities, showing high correlation with the ground truth masks (white). **(b)** On healthy control slices from the IXI dataset testing set, the model correctly produces low, diffuse activations. Colormap ranges from 0 (blue) to 12 (red).

From these visualizations, we observe that the heatmaps spatially correspond to the ground truth abnormality regions. The correspondence for the BraTS is apparent. The abnormalities in BraTS are tumors, and patches from overlapping tumors will lack the normal anatomical structure necessary for accurate location prediction. In the case of MSLUB, the presence of normal tissue surrounding small and scattered abnormal regions provides contextual cues that may allow Patch2Loc to have correct location predictions. The same applies to WMH where the long but narrow abnormal structure also provide contextual cues. Likewise, in ATLAS, if the abnormal region is very small (e.g., in the third row in Figure 9), it will go unpredicted. However, the presence of abnormalities within a patch may still cause the location prediction or increased variance in the prediction, such that Patch2Loc's abnormality score does correlate with these smaller abnormalities. That is, Patch2Loc's abnormality score is high due to in-distribution ambiguity, with large variance estimates, or out-of-distribution

patches causing error in the location prediction, and a wide range of variance estimates. These different cases can be see in the KDE plots in Figure 5.

5.3 Quantitative Results for Unsupervised Abnormality Segmentation

Performance is measured, following previous work (Behrendt et al., 2024; 2025), in terms of both average precision (area under the precision-recall curve) and the best possible Dice-coefficient (highest possible F1 score) $\lceil \text{Dice} \rceil$ per subject and then averaged over the BraTS, MSLUB, ATLAS, and WMH datasets. The results are in Table 2. Patch2Loc often matches or outperforms the best performing method, with a wide margin for $\lceil \text{Dice} \rceil$ on the WMH dataset.

	BraTS21 (T2)		MSLUB (T2)		ATLAS (T1)		WMH (T1)	
Model	$\lceil$**DICE**$\rceil$ **[%]**	**AUPRC [%]**	$\lceil$**DICE**$\rceil$ **[%]**	**AUPRC [%]**	$\lceil$**DICE**$\rceil$ **[%]**	**AUPRC [%]**	$\lceil$**DICE**$\rceil$ **[%]**	**AUPRC [%]**
Thresh (Meissen et al., 2022a)	30.26	20.27	7.65	4.23	4.66	1.71	10.32	4.72
VAE (Baur et al., 2021)	33.12 ± 1.12	25.74 ± 1.37	8.10 ± 0.18	4.48 ± 0.18	15.63 ± 0.73	11.44 ± 0.5	7.60 ± 0.28	3.86 ± 0.40
SVAE (Behrendt et al., 2022)	36.43 ± 0.36	30.3 ± 0.45	8.55 ± 0.11	4.8 ± 0.09	10.32 ± 0.53	6.84 ± 0.44	7.18 ± 0.07	2.97 ± 0.06
AE (Baur et al., 2021)	36.04 ± 1.73	28.8 ± 1.72	9.65 ± 0.97	5.71 ± 0.80	14.04 ± 0.6	10.16 ± 0.53	7.34 ± 0.08	3.43 ± 0.14
DAE (Kascenas et al., 2022)	48.82 ± 3.68	49.38 ± 4.18	7.57 ± 0.61	4.47 ± 0.69	15.95 ± 0.69	13.37 ± 0.62	12.02 ± 1.01	8.54 ± 1.02
RA (Bercea et al., 2023c)	16.75 ± 0.51	9.98 ± 0.43	3.96 ± 0.03	1.92 ± 0.04	12.21 ± 0.98	8.75 ± 0.93	6.04 ± 0.45	3.15 ± 0.31
PHANES (Bercea et al., 2023d)	28.42 ± 0.91	21.29 ± 1.06	6.11 ± 0.27	2.98 ± 0.07	17.62 ± 0.41	13.81 ± 0.48	7.55 ± 0.17	3.87 ± 0.13
FAE (Meissen et al., 2022b)	44.59 ± 2.19	43.63 ± 0.47	6.85 ± 0.65	3.85 ± 0.08	17.76 ± 0.16	13.91 ± 0.10	8.81 ± 0.38	4.77 ± 0.26
DDPM (Wyatt et al., 2022)	50.27 ± 2.67	50.61 ± 2.92	9.71 ± 1.29	6.27 ± 1.58	20.18 ± 0.58	17.77 ± 0.47	12.06 ± 0.97	8.89 ± 0.89
pDDPM (Behrendt et al., 2024)	53.61 ± 0.51	55.08 ± 0.54	12.83 ± 0.40	*10.02 ± 0.36*	19.92 ± 0.24	17.84 ± 0.10	10.13 ± 0.53	7.52 ± 0.56
cDDPM (Behrendt et al., 2025)	*56.30 ± 1.25*	**58.82 ± 1.56**	*14.04 ± 1.16*	**10.97 ± 1.17**	*24.22 ± 1.10*	**22.22 ± 1.15**	*11.59 ± 0.93*	*9.26 ± 1.07*
Patch2Loc (Ours)	**59.50 ± 1.45**	*55.40 ± 1.30*	**14.30 ± 1.40**	8.70 ± 1.20	**25.50 ± 0.73**	*22.00 ± 1.70*	**15.70 ± 1.70**	**10.10 ± 1.70**

Table 2: Comparison of the evaluated models with the best results highlighted in bold, and second best italicized. For all metrics, the mean ± standard deviation across the different folds are reported.

6 Discussion

The results showcase that Patch2Loc advances the state of the art for unsupervised abnormality segmentation in neuroimages through an intuitive approach. Its promising performance meets or exceeds the best performing benchmark cDDPM (Behrendt et al., 2025), which has a more complicated structure for training and inference due the combination of the masking autoencoder and DDPM model. In contrast, Patch2Loc is unique in terms of its dependence of local features, without global context, and it does not depend on reconstruction errors compared to the others. During test time, our method does not depend on hyperparameter such as the amount or type of noise added. For instance, DAE (Kascenas et al., 2022) has to search for the correlated noise parameter that gives the best performance. Generally, prior work requires extensive hyperparameter search during training and/or testing, especially for methods involving GANs. In contrast, Patch2Loc's hyper-parameters are optimized for the normal anatomical structure of brain images from the IXI brain slices before the models (one for T1 and T2) are evaluated on different datasets.

One limitation of Patch2Loc is its ability to infer the correct location when abnormalities are smaller than the patch size. Smaller patches deteriorate prediction performance on normal patches, due to the high similarity of patches from widely different locations.

7 Conclusion

We have introduced Patch2Loc, a novel self-supervised learning task and model tailored for lesion detection in neuroimages, and demonstrated its effectiveness for unsupervised abnormal tissue segmentation. Our approach leverages the regularities of location-specific image features to identify abnormalities in brain tissues and directly incorporates uncertainty estimates using the β-NLL framework (Seitzer et al., 2022). Unlike prior methods that focus on global features, Patch2Loc emphasizes local representations, enhancing its applicability to unsupervised abnormality segmentation in brain MRI. Our method does not need to any hyperparameter adjustment during the inference time such as the level of added noise as required by state-of-the-art methods (Behrendt et al., 2025; 2024) that rely on denoising diffusion models. This work introduces a new perspective on unsupervised abnormality segmentation in neuroimaging and lays the foundation for future research in this direction.

Large Language Models: LLM are used to polish and improve the writing of this manuscript.

REFERENCES

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

C. Baur, S. Denner, B. Wiestler, N. Navab, and S. Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study. *Medical Image Analysis*, 69: 101952, 2021.

Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*, pp. 161–169. Springer, 2019.

Finn Behrendt, Marcel Bengs, Debayan Bhattacharya, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Capturing Inter-Slice Dependencies of 3D Brain MRI-Scans for Unsupervised Anomaly Detection. *OpenReview*, April 2022. URL https://openreview.net/forum?id=db8wDgKH4p4.

Finn Behrendt, Debayan Bhattacharya, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Patched diffusion models for unsupervised anomaly detection in brain mri. In *Medical Imaging with Deep Learning*, pp. 1019–1032. PMLR, 2024.

Finn Behrendt, Debayan Bhattacharya, Robin Mieling, Lennart Maack, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Guided reconstruction with conditioned diffusion models for unsupervised anomaly detection in brain mris. *Computers in Biology and Medicine*, 186:109660, 2025.

Cosmin Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia Schnabel. Evaluating normative learning in generative ai for robust medical anomaly detection. 2023a.

Cosmin I. Bercea, Michael Neumayr, Daniel Rueckert, and Julia A Schnabel. Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*, 2023b. URL https://openreview.net/forum?id=kTpafpXrqa.

Cosmin I. Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A. Schnabel. Generalizing Unsupervised Anomaly Detection: Towards Unbiased Pathology Screening, April 2023c. URL https://openreview.net/forum?id=8ojx-Ld3yjR. [Online; accessed 2. Jun. 2025].

Cosmin I Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A Schnabel. Reversing the abnormal: Pseudo-healthy generative networks for anomaly detection. *arXiv preprint arXiv:2303.08452*, 2023d.

Biomedical Image Analysis Group, Imperial College London. Ixi dataset, 2015. URL https://brain-development.org/ixi-dataset/. This work is licensed under the CC-BY-SA 3.0. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/3.0/legalcode.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

L. Busby, J. Courtier, and C. Glastonbury. Bias in radiology: the how and why of misses and misinterpretations. *Radiographics*, 38:236, 2018.

Xiaoran Chen and Ender Konukoglu. Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. *arXiv preprint arXiv:1806.04972*, 2018.

Xiaoran Chen, Suhang You, Kerem Can Tezcan, and Ender Konukoglu. Unsupervised lesion detection via image restoration with a normative prior. *Medical Image Analysis*, 64:101713, 2020.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

M. Hagens, J. Burggraaff, I. Kilsdonk, S. Ruggieri, S. Collorone, R. Cortese, N. Cawley, E. Sbardella, M. Andelova, M. Amann, et al. Impact of 3 tesla mri on interobserver agreement in clinically isolated syndrome: a magnims multicentre study. *Multiple Sclerosis Journal*, 25: 352–360, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

C. Josephson, J. Bhattacharya, C. Counsell, V. Papanastassiou, V. Ritchie, R. Roberts, R. Sellar, C. Warlow, and R. Salman. Seizure risk with avm treatment or conservative management: Prospective, population-based study. *Neurology*, 79:500–507, 2012.

Antanas Kascenas, Nicolas Pugeault, and Alison Q O'Neil. Denoising autoencoders for unsupervised anomaly detection in brain mri. In *International Conference on Medical Imaging with Deep Learning*, pp. 653–664. PMLR, 2022.

Antanas Kascenas, Pedro Sanchez, Patrick Schrempf, Chaoyang Wang, William Clackett, Shadia S. Mikhael, Jeremy P. Voisey, Keith Goatman, Alexander Weir, Nicolas Pugeault, Sotirios A. Tsaftaris, and Alison Q. O'Neil. The role of noise in denoising models for anomaly detection in medical images. *Med. Image Anal.*, 90:102963, December 2023. ISSN 1361-8415. doi: 10.1016/j.media.2023.102963.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*, 2014. URL `https://www.semanticscholar.org/paper/Adam%3A-A-Method-for-Stochastic-Optimization-Kingma-Ba/a6cb366736791bcccc5c8639de5a8f9636bf87e8`.

Hugo J. Kuijf, J. Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M. Jorge Cardoso, Adria Casamitjana, D. Louis Collins, Mahsa Dadar, Achilleas Georgiou, Mohsen Ghafoorian, Dakai Jin, April Khademi, Jesse Knight, Hongwei Li, Xavier Llado, Miguel Luna, Qaiser Mahmood, Richard McKinley, Alireza Mehrtash, Sebastien Ourselin, Bo-Yong Park, Hyunjin Park, Sang Hyun Park, Simon Pezold, Elodie Puybareau, Leticia Rittner, Carole H. Sudre, Sergi Valverde, Veronica Vilaplana, Roland Wiest, Yongchao Xu, Ziyue Xu, Guodong Zeng, Jianguo Zhang, Guoyan Zheng, Christopher Chen, Wiesje van der Flier, Frederik Barkhof, Max A. Viergever, and Geert Jan Biessels. Standardized Assessment of Automatic Segmentation of White Matter Hyperintensities and Results of the WMH Segmentation Challenge. *IEEE Trans. Med. Imaging*, 38(11):2556–2568, November 2019. ISSN 1558-254X. doi: 10.1109/TMI.2019.2905770.

Žiga Lesjak, Alfiia Galimzianova, Aleš Koren, Matej Lukin, Franjo Pernuš, Boštjan Likar, and Žiga Špiclin. A novel public mr image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. *Neuroinformatics*, 16:51–63, 2018.

Ziyun Liang, Harry Anthony, Felix Wagner, and Konstantinos Kamnitsas. Modality cycles with masked conditional diffusion for unsupervised anomaly segmentation in mri, 2023.

Sook-Lei Liew, Bethany P. Lo, Miranda R. Donnelly, Artemis Zavaliangos-Petropulu, Jessica N. Jeong, Giuseppe Barisano, Alexandre Hutton, Julia P. Simon, Julia M. Juliano, Anisha Suri, Zhizhuo Wang, Aisha Abdullah, Jun Kim, Tyler Ard, Nerisa Banaj, Michael R. Borich, Lara A. Boyd, Amy Brodtmann, Cathrin M. Buetefisch, Lei Cao, Jessica M. Cassidy, Valentina Ciullo, Adriana B. Conforto, Steven C. Cramer, Rosalia Dacosta-Aguayo, Ezequiel de la Rosa, Martin Domin, Adrienne N. Dula, Wuwei Feng, Alexandre R. Franco, Fatemeh Geranmayeh, Alexandre Gramfort, Chris M. Gregory, Colleen A. Hanlon, Brenton G. Hordacre, Steven A. Kautz, Mohamed Salah Khlif, Hosung Kim, Jan S. Kirschke, Jingchun Liu, Martin Lotze, Bradley J. MacIntosh, Maria Mataró, Feroze B. Mohamed, Jan E. Nordvik, Gilsoon Park, Amy Pienta, Fabrizio Piras, Shane M. Redman, Kate P. Revill, Mauricio Reyes, Andrew D. Robertson, Na Jin Seo, Surjo R. Soekadar, Gianfranco Spalletta, Alison Sweet, Maria Telenczuk, Gregory Thielman, Lars T. Westlye, Carolee J. Winstein, George F. Wittenberg, Kristin A. Wong, and Chunshui Yu. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Sci. Data*, 9(320):1–12, June 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01401-7.

Felix Meissen, Georgios Kaissis, and Daniel Rueckert. Challenging Current Semi-supervised Anomaly Segmentation Methods for Brain MRI. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pp. 63–74. Springer, Cham, Switzerland, July 2022a. ISBN 978-3-031-08999-2. doi: 10.1007/978-3-031-08999-2_5.

Felix Meissen, Johannes Paetzold, Georgios Kaissis, and Daniel Rueckert. Unsupervised Anomaly Localization with Structural Feature-Autoencoders. *arXiv*, August 2022b. doi: 10.1007/978-3-031-33842-7_2.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34 (10):1993–2024, 2014.

Leonard Merewitz and Jonathan H Sunshine. A portrait of pediatric radiologists in the united states. *American Journal of Roentgenology*, 186(1):12–22, 2006.

David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pp. 55–60. IEEE, 1994.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pp. 69–84. Springer, 2016.

L. G. Nyul, J. K. Udupa, and Xuan Zhang. New variants of a method of MRI scale standardization. *IEEE Trans. Med. Imaging*, 19(2):143–150, February 2000. doi: 10.1109/42.836373.

W. Pinaya, P. Tudosiu, R. Gray, G. Rees, P. Nachev, S. Ourselin, and M. Cardoso. Unsupervised brain imaging 3d anomaly detection and segmentation with transformers. *Medical Image Analysis*, 79:102475, 2022a.

Walter HL Pinaya, Mark S Graham, Robert Gray, Pedro F Da Costa, Petru-Daniel Tudosiu, Paul Wright, Yee H Mah, Andrew D MacKinnon, James T Teo, Rolf Jager, et al. Fast unsupervised brain anomaly detection and segmentation with diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 705–714. Springer, 2022b.

Torsten Rohlfing, Natalie M. Zahr, Edith V. Sullivan, and Adolf Pfefferbaum. The SRI24 multichannel atlas of normal adult human brain structure. *Hum. Brain Mapp.*, 31(5):798, December 2009. doi: 10.1002/hbm.20906.

Kazuki Sato, Kenta Hama, Takashi Matsubara, and Kuniaki Uehara. Predictable uncertainty-aware unsupervised deep anomaly segmentation. In *2019 International Joint Conference on Neural Networks (ijcnn)*, pp. 1–7. IEEE, 2019.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aPOpXlnV1T.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18158–18172. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d2dc6368837861b42020ee72b0896182-Paper.pdf.

K. Van Hespen, J. Zwanenburg, J. Dankbaar, M. Geerlings, J. Hendrikse, and H. Kuijf. An anomaly detection approach to identify chronic brain infarcts on mri. *Scientific Reports*, 11:1–10, 2021.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris G Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 650–656, 2022.

David Zimmerer, Simon AA Kohl, Jens Petersen, Fabian Isensee, and Klaus H Maier-Hein. Context-encoding variational autoencoder for unsupervised anomaly detection. *arXiv preprint arXiv:1812.05941*, 2018.

A Appendix

A.1 Related Work

Unsupervised abnormality segmentation in neuroimaging has benefited from improving machine learning models ranging from autoencoders (AE), variational autoencoders (VAE) (Baur et al., 2021), generative adversarial neural networks (GANs) or discriminators as in GANs (Goodfellow et al., 2014), transformers, and diffusion models. Nonetheless, a notable baseline is based on a simple threshold on the image histogram (Meissen et al., 2022a), which can exceed the performance of baseline methods for certain modalities.

Many approaches leverage AE or VAE in novel ways. An early work by Zimmerer et al. (2018) introduced *ceVAE*, which combines variational and context autoencoders to compute abnormality scores from density and reconstruction error. Another work by Sato et al. (2019) uses a VAE with the abnormality score defined as the reconstruction error divided by the estimated variance. Rather than relying on squared error for training, another work by Meissen et al. (2022b) uses an AE architecture that reconstructs features obtained from a pretrained encoder using the Structural Similarity Index Measure (SSIM) as the loss function. SVAE by Behrendt et al. (2022) uses a VAE with transformers to capture the inter-slice dependencies and showed it can improve the results compared to 2D vanilla VAE. The RA method by Bercea et al. (2023c) uses an VAE with a cyclic loss and use the reconstruction error as abnormality score.

Previous studies (Baur et al., 2021; Pinaya et al., 2022a) noted that AE and VAE methods suffer from blurry reconstructed images, which hinders their performance. Incorporating a discriminator as in a GAN can improve the reconstruction quality. However, methods incorporating GAN-losses suffer from training instability due the competition between the discriminator and decoder (Brock et al., 2018). Nonetheless, the work by Chen & Konukoglu (2018) proposed an adversarial autoencoder (AAE) that enforces a prior on the latent space and a cyclic objective for encoder consistency such that an image and its reconstruction are similar in the latent space. AnoVAGAN by Baur et al. (2019) is a variational autoencoder with spatially organized latent codes and a GAN loss.

Another approach is to iteratively restore an image to better match the normal data distribution, using the number of restoration steps as an abnormality score (Chen et al., 2020). PHANEs by Bercea et al. (2023d) uses a model to restore part of an MRI slice flagged by the RA method (Bercea et al., 2023c) to mitigate false positives within the flagged region.

One prior work by Van Hespen et al. (2021) used a patch-based auto-encoder with a cycle consistency term and a discriminator to distinguish between the original image and its reconstruction,

testing it on specific abnormal tissues.[1] We identified it as the only approach leveraging local features for unsupervised abnormality segmentation in brain MRI. However, it suffers from instability due to adversarial training and difficulty in balancing multiple loss terms, which is challenging in unsupervised setting. The method was applied to detect infarcts, but it was not applied to abnormality segmentation and was not compared against any existing work.

While the aforementioned works advanced unsupervised abnormality detection for neuroimages, all are outperformed by a significant margin by using a denoising autoencoder (DAE) (Kascenas et al., 2022), which learns to remove correlated noise added to the input images during training. DAE outperformed GAN and VAE approaches, achieving higher Dice score and average precision.

A promising work by Pinaya et al. (2022a) introduced a combined vector-quantized variational autoencoder to learn spatial latent representation for brain slices and then used it to train an autoregressive transformer that operates on patches of the latent representation using different raster orders. Unfortunately, the results are not comparable as they are limited to the FLAIR modality, with 15,000 normal scans sourced from the UK Biobank (UKB) dataset, which is not freely available, and the method was not tested with other modalities or datatsets. It should be noted that the IXI dataset used for training in our benchmark does not have FLAIR modality.

While not used for unsupervised abnormality segmentation, one prior work by Taleb et al. (2020) uses self-supervised learning with a task that resembles Patch2Loc's location prediction task. It is a 3D version of a previous context-dependent self-supervised task of predicting a patch's relative location with respect to a context patch as a classification problem (Doersch et al., 2015). Specifically, the task consists of predicting the discrete relative location of a 3D patch among the possible locations in a $3 \times 3 \times 3$ grid surrounding a center patch that is provided as input. Essentially, the goal is to infer the relative location of one patch with respect to center patch as context. The performance for unsupervised abnormality segmentation was not benchmarked as the task served only as a pretext task to learn latent representation via a backbone network that improves subsequent supervised segmentation performance (Taleb et al., 2020). Patch2Loc is distinguished in that its training directly informs the abnormality score. Additionally, Patch2Loc predicts the continuous location of a patch without the need of a center patch as context, using only the slice index as sufficient context for registered brain images.

A.1.1 Denoising Diffusion Probabilistic Models

Many of the top performing methods for unsupervised abnormality detection use denoising diffusion probabilistic models (DDPM). Unless stated otherwise, all methods use the absolute reconstruction error between input and denoised ones as the use the abnormality score. One of the first works to use DDPM (Wyatt et al., 2022) also proposed learning to denoise simplex noise instead of Gaussian noise to enhance performance. While this empirically works, how this matches the fundamental assumptions of diffusion processes is not clear. Specifically, simplex noise is procedurally generated, and is not described by a random process. In contrast, a diffusion model's forward process is described by a Markov chain, with a Markov transition kernel (Sohl-Dickstein et al., 2015), typically described by adding Gaussian noise, but the Markov chain can also be described by a Bernoulli transition kernel (Sohl-Dickstein et al., 2015). Without a known Markov transition kernel, the derivation of the DDPM formulation (Ho et al., 2020) may not be applicable to simplex noise, since the Gaussian assumption is exploited to define the forward process posterior mean. Nonetheless, ignoring the validity, the DDPM formulation essentially defines a training regime for denoising across different signal-noise levels, where longer times correspond to higher noise regimes. During inference, denoising is performed from the initially high noise regime, then new simplex noise is applied at decreasing noise levels, and the process is repeated.

Another work by Pinaya et al. (2022b) trained diffusion models on the learned spatial latent features and during inference time the abnormal features are inpainted with normal ones. Then the reconstructed image is obtained using the denoised spatial latent features that are fed to a decoder. Although their work showed an impressive performance for head CT, for brain MRI, their work did not exceed their earlier work (Pinaya et al., 2022a). This was extended by another work by

[1] We note that using a variational autoencoder and providing location information as input could lead to a complementary framework to Patch2Loc, where the patch reconstruction error and uncertainty regarding the latent embedding could form an abnormality score.

Liang et al. (2023) that leveraged a diffusion model for cyclic translations between different modalities and implemented a conditional model, akin to a restoration-based approach. While the model shows superior performance compared to all other techniques mentioned above it requires different modalities.

Similarly, another work by Bercea et al. (2023b) introduced an iterative inpainting technique to address the noise paradox in order to mitigate the false positives in a high noise regime. The approach initially estimates mask for abnormal tissues based on reconstruction error in the high noise regime. Then an iterative method is applied to inpaint regions of the mask using information outside the mask in a lower noise regime.

The patched diffusion model (pDDPM) by Behrendt et al. (2024) performs the noising and denoising within a patch of the whole slice. That is the rest of the slice gives context for the DDPM of the noised patch. Versions of each slice with patches at different locations are used to identify abnormalities within the slice. In a follow-up work (Behrendt et al., 2025) a conditional DDPM (cDDPM) is used for denoising. The conditioning signal is the latent embedding from a masked auto-encoder (MAE) (He et al., 2022) pretrained on normal MRI slices, which is further fine-tuned during training. This conditioning gives the model global perspective of structure while performing the denoising.

A.2 More Figures

This section contains additional figures referenced in the main body.

Figure 5 show kernel density estimates of the log-error and log-variance of normal and abnormal patches from the test set of IXI (T1 and T2 modalities), ATLAS, MSLUB, and WMH datasets.

Figures 6 7, 9, 8, and 10 show example neuroimages, Patch2Loc's heatmap, and the ground truth for representative slices from IXI, BraTS, ATLAS, and MSLUB datasets, respectively.

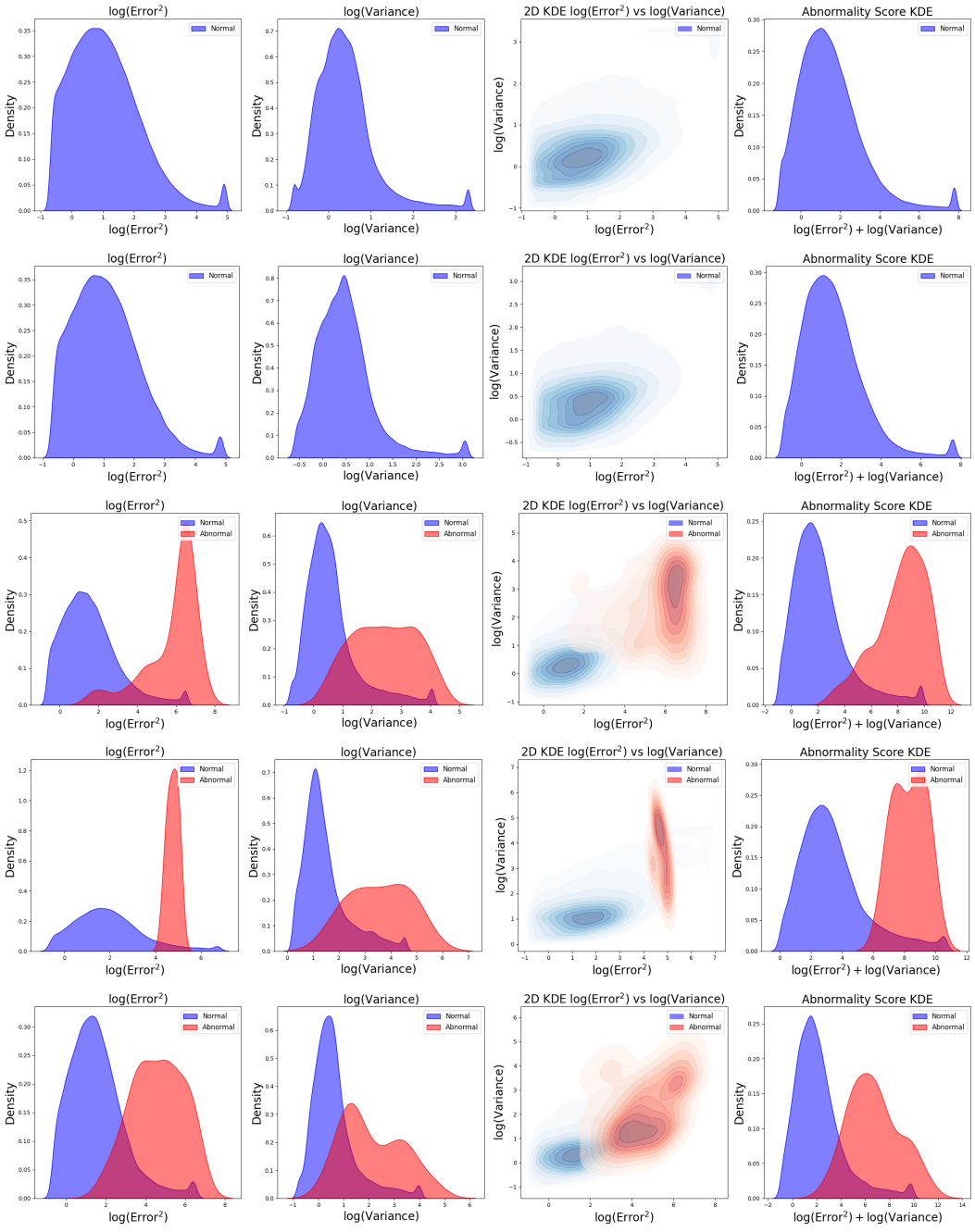


Figure 5: From left to right: 1D KDE for log-error, and log-variance, 2D KDE for log-error and log-variance, and 1D KDE for the abnormality score for normal patches (blue) and abnormal patches (red). Top to bottom: test set IXI (T1), test set IXI (T2), ATLAS, MSLUB, and WMH datasets.

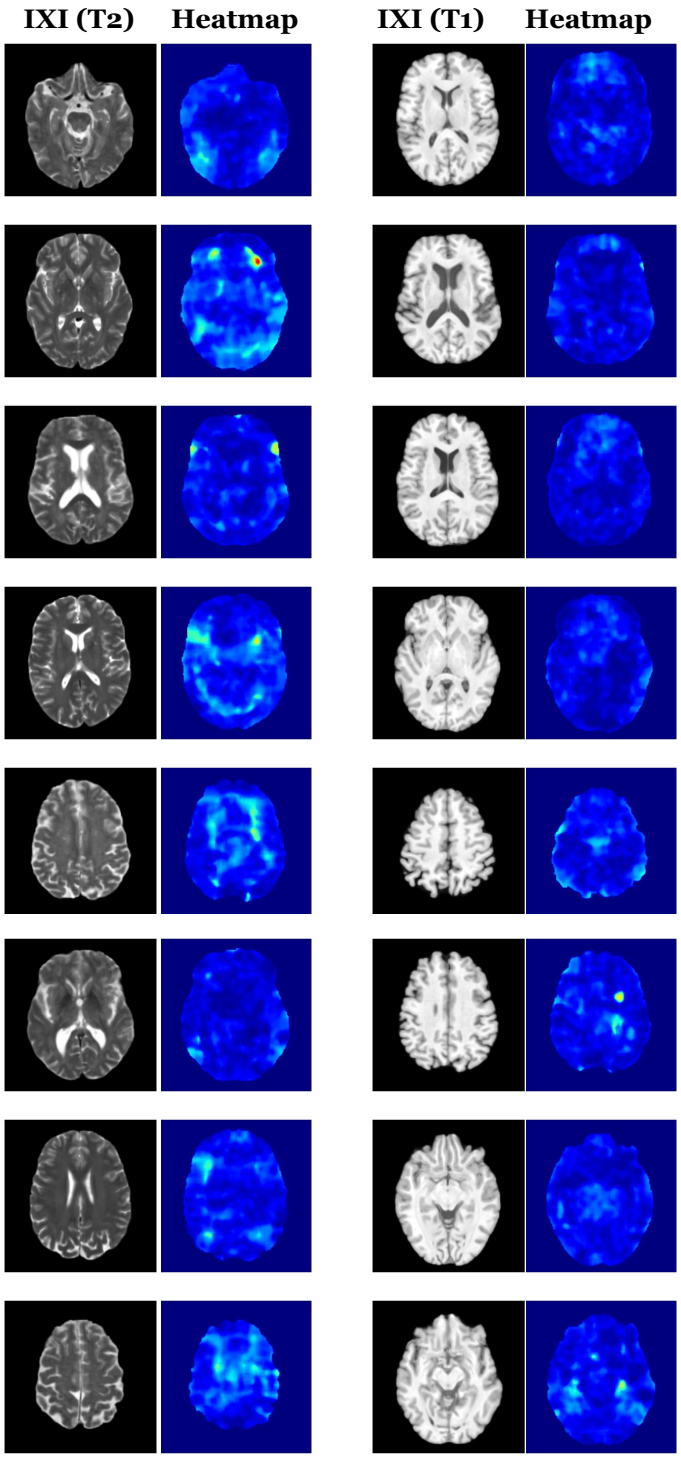


Figure 6: Visualization of slices from IXI (T1 and T2). Colormap ranges from 0 (blue) to 12 (red).

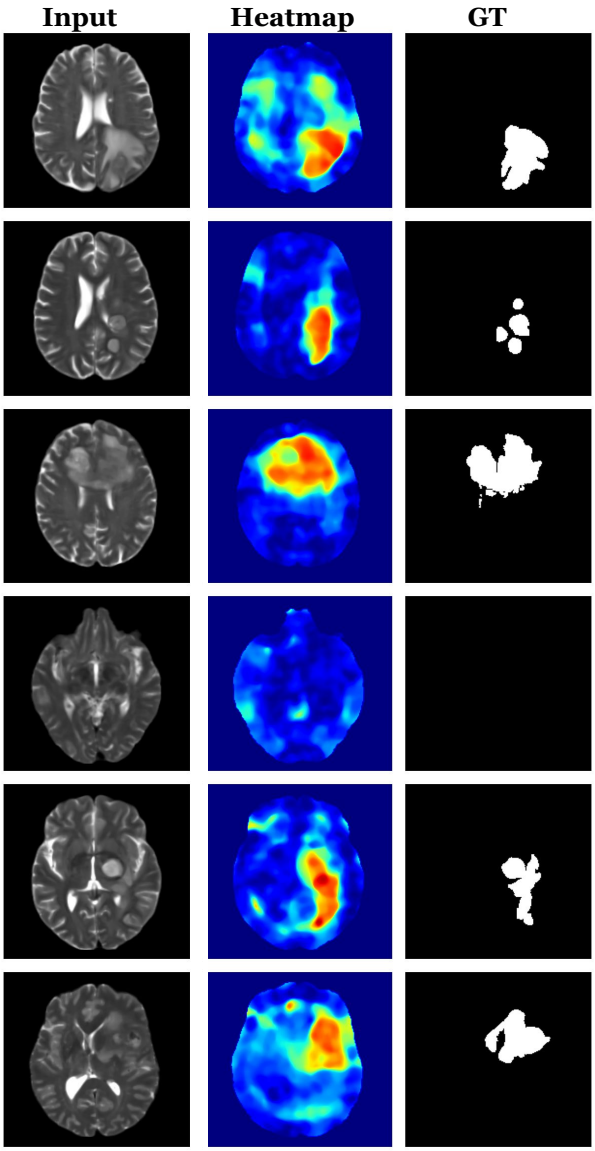


Figure 7: Visualization of slices from BraTS. From left to right: the input slice, the heatmap (ranges from 0 (blue) to 12 (red)), and the ground truth for abnormal tissues.

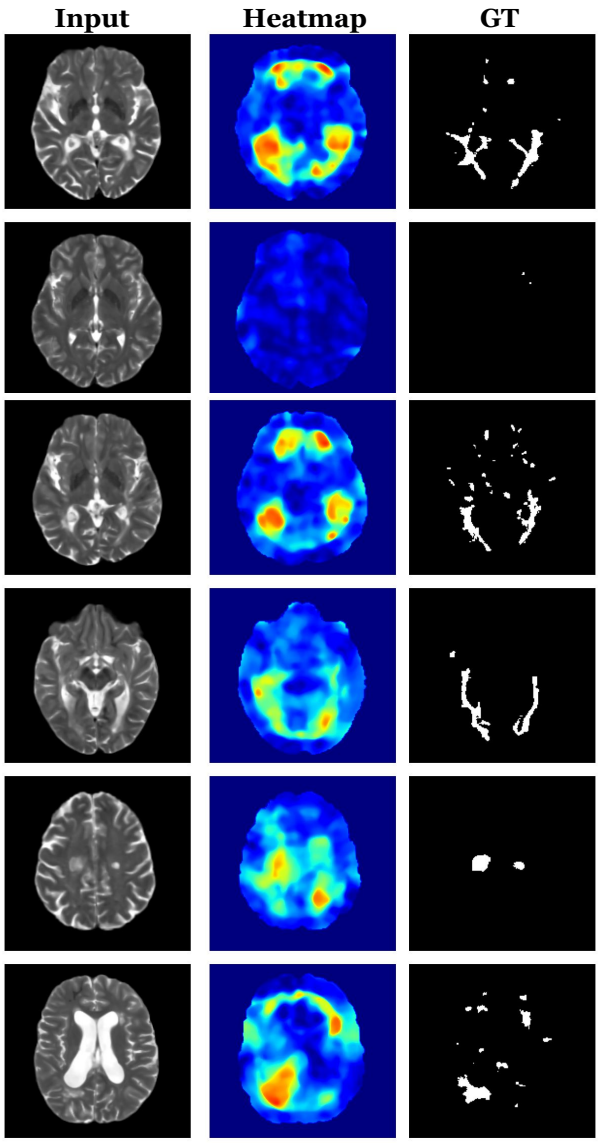


Figure 8: Visualization of slices from MSLUB. From left to right: the input slice, the heatmap (ranges from 0 (blue) to 12 (red)), and the ground truth for abnormal tissues.

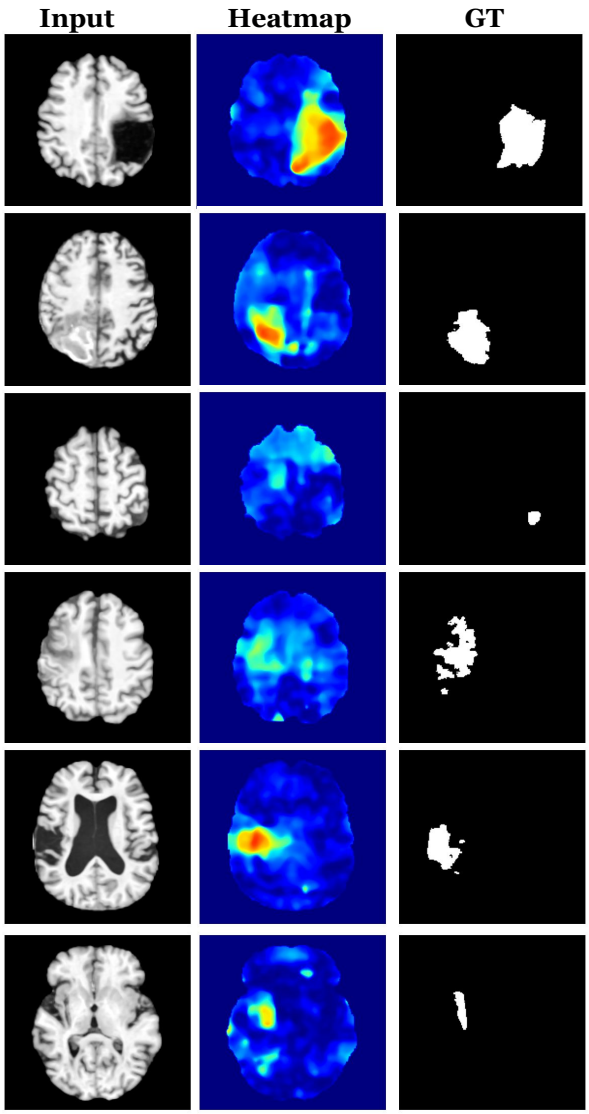


Figure 9: Visualization of slices from ATLAS. From left to right: the input slice, the heatmap (ranges from 0 (blue) to 12 (red)), and the ground truth for abnormal tissues.

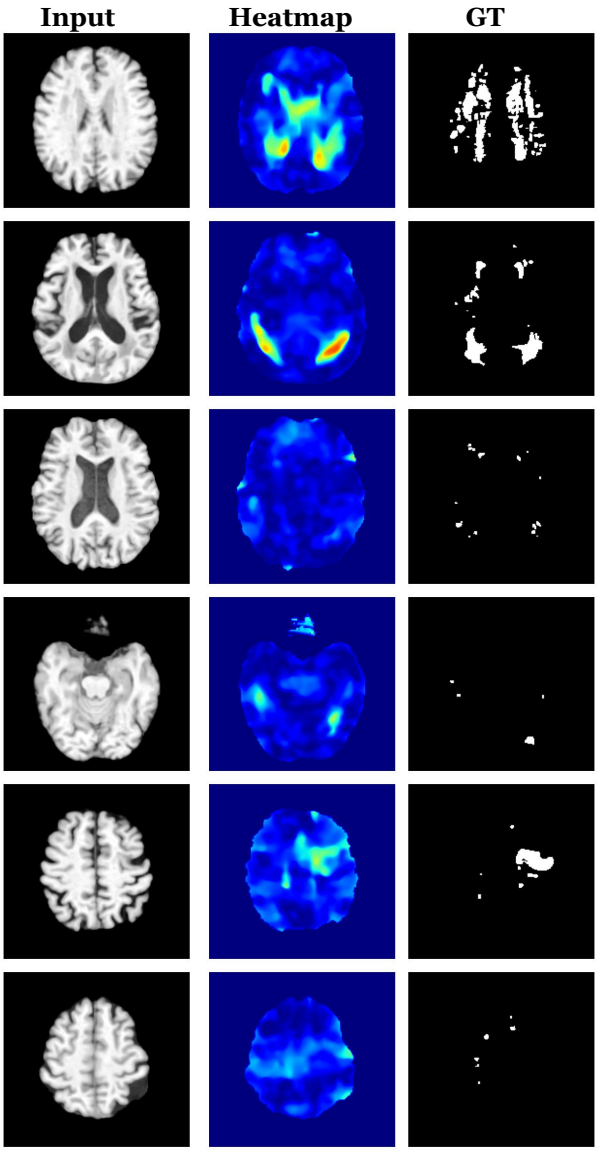


Figure 10: Visualization of slices from WMH. From left to right: the input slice, the heatmap (ranges from 0 (blue) to 12 (red)), and the ground truth for abnormal tissues.