

Spatially Context-Aware Transformers Facilitate Modeling-Based Anomaly Detection of Subtle Lesions in Brain MRI Images

Johannes Schwarz ³, Lena Will ^{3,4}, Jörg Wellmer ³, Axel Mosig ^{*} ^{1,2}

1 Ruhr-University Bochum, Center for Protein Diagnostics, Bochum, Germany

2 Ruhr-University Bochum, Faculty of Biology and Biotechnology, Bioinformatics Group, Bochum, Germany

3 Ruhr-Epileptology, Department of Neurology, University Hospital Knappschaftskrankenhaus Bochum, Ruhr-University Bochum, Bochum, Germany

4 Ruhr-Epileptology, Department of Radiology and Neuroradiology, University Hospital Knappschaftskrankenhaus Bochum, Ruhr-University Bochum, Bochum, Germany

Abstract

The detection of small and subtle lesions in high-resolution 3D volumes is a highly relevant, yet far from solved task in biomedical imaging. We here address a specific task in detecting certain types of epileptogenic lesions through our novel semi-supervised *spatially context aware transformer* (*SpyCAT*) approach to anomaly detection. *SpyCAT* is modeling-based in the sense that it builds on specific assumptions that constitute what is normal and what constitutes relevant deviations from normality. We explicitly use these assumptions to justify the inductive bias of our anomaly detection approach. The resulting *SpyCAT* system is patch-based and uses a transformer architecture to process discrete tokens obtained from a vector-quantized variational autoencoder, which produces counterfactual patches through full 3D convolutions of each patch. We evaluate our approach on the grounds of point-annotations of two subtypes of epileptogenic lesions, using validation measures that build on the *Metrics Reloaded* framework, showing that *SpyCAT* can reliably identify and localize the lesion types under consideration, and outperforms state-of-the-art reference methods. Our code is available at <https://github.com/johannesSX/SpyCAT>.

Keywords

magnetic resonance imaging, , anomaly detection, , semi-supervised learning, , spatially contextualized patch transformers, , voxel and image-level detection

Article informations

<https://doi.org/10.59275/j.melba.2026-a92b>

©2026 Schwarz, Will, Wellmer and Axel. License: CC-BY 4.0

Received: 2025-07-25, Published 2026-10-18

Corresponding author: axel.mosig@ruhr-uni-bochum.de



1. Introduction

Anomaly detection approaches have been widely investigated for the detection of brain lesions in MRI images (Simarro et al., 2020; Bowles et al., 2017; Kascenas et al., 2022; Meissen et al., 2023; Baur et al., 2021). The vast majority of these approaches have been validated on relatively large lesions, mostly brain tumors, that are correspondingly easy to detect (Luo et al., 2023; Simarro et al., 2020; Wyatt et al., 2022). However, in many cases, neurological disorders are associated with brain lesions that are much smaller and subtler than the tumors present in the datasets commonly used to assess brain anomaly detection approaches.

Using anomaly detection for identifying epileptogenic

lesions is particularly challenging. In a cohort of 2,740 patients, Wellmer et al. (2013) identified a spectrum of 16 different entities of epilepsy-associated lesions. Therapy-relevant lesions are in many cases 5 – 7 mm or even smaller in size, and detection of the full spectrum of epileptogenic lesions requires a broad bandwidth of high resolution MRI sequences. The study by Wellmer et al. (2013) involved T1-weighted, T2-weighted, FLAIR and SWI sequences. The challenge of epileptogenic lesion detection has also been recognized in the literature (Spitzer et al., 2022), and the question whether subtle lesions can be detected by anomaly detection approaches must be considered open. In this contribution, we propose the *spatially context aware transformer* (*SpyCAT*) as a novel semi-supervised anomaly de-

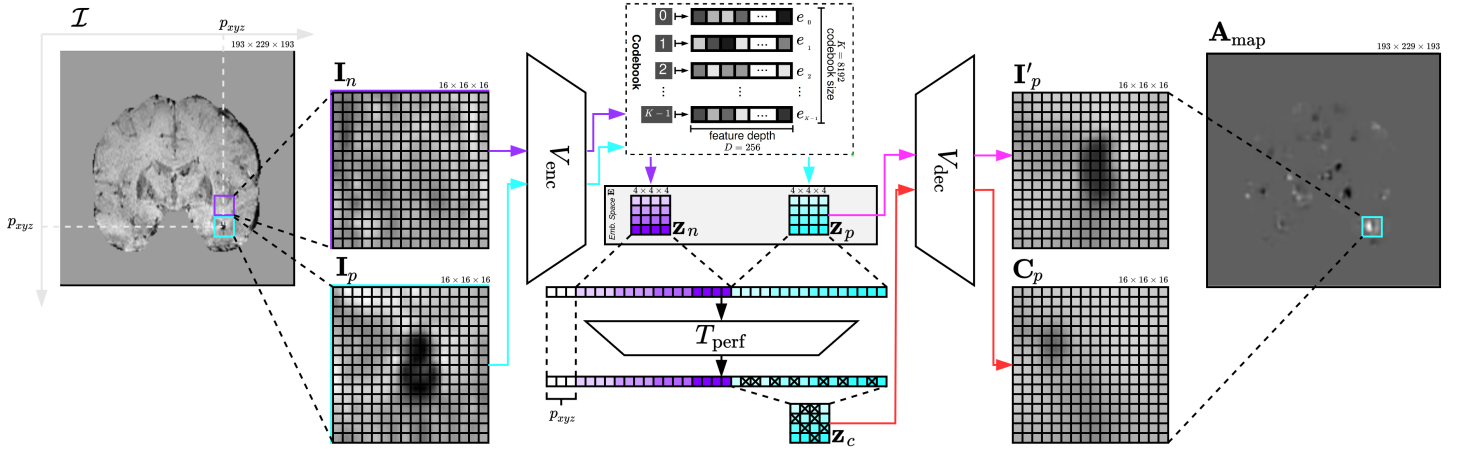


Figure 1: Schematic model architecture for simplicity with 2D images (sampled from coronal view of the MRI). The VQ-VAE (consists of V_{enc} and V_{dec}) receives an input patch $\mathbf{I}_p$, a neighborhood patch $\mathbf{I}_n$ and generates an embedding space $\mathbf{z}_p$ and $\mathbf{z}_n$. By autoregression, the transformer T_{perf} creates a counterfactual (healed) embedding space $\mathbf{z}_c$, given $\mathbf{z}_n$ and a 3D position coordinate p , which describes where the patch is located in the MRI. Using the decoder V_{dec} , the original embedding space $\mathbf{z}_p$ and the counterfactual embedding space $\mathbf{z}_c$ become $\mathbf{I}'_p$ and $\mathbf{C}_p$. Subtraction between $\mathbf{I}'_p$ and $\mathbf{C}_p$ generates a patch anomaly map and these anomaly maps can be combined to an anomaly map $\mathbf{A}_{map}$ by rasterization.

tection approach that addresses this question in the context of localizing certain subtypes of epilepsy-associated brain lesions on MRI images.

As illustrated in Figure 1, our approach is based on a set of explicit assumptions which state that small and subtle lesions can be identified as anomalies that deviate in contrast to the morphological context in their non-lesional, local spatial surrounding. We translate these explicit assumptions into a transformer-based approach that operates on feature embeddings obtained from a vector-quantized variational autoencoder (VQ-VAE). The feature embeddings are computed on small patches whose size reflects the expected small size of the lesions to be detected. Spatial context is provided by feature embeddings of a neighboring patch, so that a transformer network can be trained and applied to a concatenation of a repertoire of spatial context, a position encoding, and the patch of interest, in which a lesion is to be detected. This yields a transformed patch embedding, whose deviation from an expected normal, lesion-free patch can be measured against a naive patch reconstruction.

We evaluate and compare our method on three datasets against four state-of-the-art anomaly detection approaches: the β -variational autoencoder (Loizillon et al., 2024), f-AnoGAN (Schlegl et al., 2019; Simarro et al., 2020), PatchCore (Roth et al., 2021), and DenoisingAE (Kascenas et al., 2023). Our first dataset is an in-house collection of MRI studies with two subtypes of epilepsy-associated lesions (for simplicity, epilepsy-associated or epileptogenic lesions are summarized under the acronym EL). As second and third datasets, we assess detection performance on the

well-established BraTS 2021 dataset (Menze et al., 2015; Bakas et al., 2017, 2018) for detecting primary brain tumors (gliomas) and on the BraTS-METS 2023 dataset (Moawad et al., 2023, 2024) for detecting brain metastases.

2. Related Work

Anomaly detection has been widely investigated in medical imaging (Cai et al., 2025). Many state-of-the-art anomaly detection approaches have been developed outside the medical domain around the MVTEC AD dataset (Bergmann et al., 2019) which consists of images of industrial parts. Numerous anomaly detection approaches have been proposed that deliver strong performance (Lee et al., 2022; Defard et al., 2020; Roth et al., 2021) on this particular dataset. Yet, these approaches are tailored towards 2D applications and raise the question whether anomalies of non-medical objects in the MVTEC dataset are set apart from normal data in the same manner as anomalies in medical image data. In the context of MRI images, anomaly detection approaches are often evaluated on the BraTS tumor dataset (Menze et al., 2015; Bakas et al., 2017, 2018; Schwarz et al., 2024) in which the anomalies cover a relatively large fraction of the image volume and are correspondingly clearly visible and can be identified as global anomalies utilizing 3D approaches (Simarro et al., 2020; Bowles et al., 2017; Iqbal et al., 2023; Luo et al., 2023). Many approaches work on the level of 2D slices, where the BR35H dataset (Hamada, 2020) is commonly used to investigate detection performance (Guo et al., 2024; Wyatt et al., 2022; Behrendt et al., 2023).

In some contexts, for example when detecting epilepsy-associated lesions (Walger et al., 2023) or multiple sclerosis lesions (Carass et al., 2017), the lesions to be detected are smaller and more subtle than brain tumors. While multiple sclerosis lesion detection is often guided by longitudinal data (Carass et al., 2017), epileptogenic lesions usually need to be detected in single MRI images. For this purpose, in particular for detecting focal cortical dysplasias, supervised methods have been used more predominantly (Walger et al., 2023), while anomaly detection has so far taken no relevant role. Corresponding supervised classifiers have, for example, been proposed by Gill et al. (2021) and Spitzer et al. (2022). To find very small lesions, a focus hypothesis using EEG can be helpful and improves the result of the classifier (Gill et al., 2021).

In many of these studies, the lesion data are usually obtained from the same scanner and were recorded in the same hospital using one and the same protocol. As Bottani et al. (2023) have identified, the generalizability to clinical routine data is severely limited. A few approaches utilize collections of lesion-free reference data such as the IXI dataset¹ to learn the patterns of *normal* datasets, and then identify deviations from those as lesional anomalies (Wyatt et al., 2022). However, these approaches rely either on anomaly detection in 2D sections or lack multicentric validation. On MRI data, anomaly detection is often carried out using supervised (Gill et al., 2021; Spitzer et al., 2022) and semi-supervised (Simarro et al., 2020; Iqbal et al., 2023; Luo et al., 2023; Marimont and Tarroni, 2020) methods. Annotation in 3D space being very time demanding, semi-supervised approaches are of high relevance, since training requires only *normal* MRIs, while annotated lesional MRIs are only required for validation.

Small and subtle lesions impose additional challenges to anomaly detection, since the small local deviations may be overshadowed by non-lesional morphological variances of a specific individual brain. This can in general be addressed by patch-based learning, where the brain image volume is decomposed into smaller patches. Recently, Behrendt et al. (2023) used this approach to reliably reconstruct normal brain MRI patches, which can be contrasted with matching patches in a given MRI to reveal lesions. In a sense, patch decomposition adds prior knowledge to anomaly detection, as the patch size is usually chosen according to the expected size of the lesions to be detected. This consideration is in line with the modeling approach suggested by Baur et al. (2021), in which the detection of abnormalities involves modeling assumptions what variances in the data are considered normal.

Vector-quantized variational autoencoders (VQ-VAEs) (Oord et al., 2017) have been utilized in different contexts

for anomaly detection tasks (Malé et al., 2024). It is a somewhat natural idea to use discrete tokens obtained from VQ-VAEs as tokens for a transformer architecture, which has emerged independently in the literature in several places (Wang et al., 2023; Lu et al., 2023; Patel et al., 2023; Fayyaz et al., 2022)

Recently, diffusion-based models exhibited strong performance in medical image classification (Favero et al., 2025) and related tasks such as the generation of counterfactual images (Fathi et al., 2024). As recently surveyed in the comprehensive survey by Bercea et al. (2025), diffusion approaches exhibit strong performance across a variety of tasks in medical image analysis, including brain lesion detection. For the specific task of detecting small brain lesions, the review (Bercea et al., 2025) identifies a specific weakness of the recent patch-DDPM (Behrendt et al., 2023) to miss small lesions. As the review by Bercea et al. (2025) convincingly concludes, the limitation of current diffusion model approaches may be that they are implemented in 2D, and overcoming the challenges of high memory requirements and long sampling times when implementing 3D diffusion models may boost their performance for detecting small lesions in 3D MR or CT scans.

3. Approach

When considering anomaly detection approaches for medical imaging data, it is inevitable to realize that anomalies can occur in many types: Some anomalies can be sharply localized point divergences, while others can be diffuse morphological alterations or even global structural deviations compared to images of normal, healthy or wildtype entities. Thus, when applying anomaly detection in a particular setting, it is inevitable to raise the question what types of anomalies are relevant. In other words, one needs to constitute a notion of normality, and decide what types of deviations from normality represent relevant anomalies. In the lack of an established term for this constituting process, we refer to it as *contextualizing normality*.

Our *SpyCAT* algorithm contextualizes normality through a targeted modeling approach. This approach is based on explicitly phrased assumptions about how epileptogenic lesions are manifested in brain MRIs in contrast to non-lesional MRIs. These modeling assumptions are then used to guide the inductive bias of the deep learning architecture to identify deviations from those variances that are considered normal.

Informally speaking, abnormal, i.e. lesional brain MRIs, are distinguished from non-lesional MRIs by regions that involve only a small fraction of the overall brain volume. This assumption calls for presupposing a *local* notion of abnormality. As a first step towards translating this into a formal deep learning approach, we explicitly postulate the

1. <https://brain-development.org/ixi-dataset/>

following assumptions about how epileptogenic lesions are distinguished in contrast to normal brain MRIs:

- (A1) There is an upper limit of the brain volume that an EL can typically cover.
- (A2) ELs are spatially surrounded by *normal* brain regions, so that the ELs deviate from the morphology that is to be expected from surrounding brain regions.
- (A3) The MRI morphology of specific localized brain regions in *normal* MRIs is predictable from surrounding brain regions.
- (A4) Brain MRIs can be registered into a coordinate system where the same brain regions across different brains are localized at similar coordinates.

The need for contextualizing normality can also be viewed from the perspective of machine learning theory. For supervised learning, machine learning theory is famously built on Valiant's concept of *probably approximate correct* (PAC) learnability. PAC learning deals with a hypothesis space as a set of functions that map from the domain of data points to a label space, and such space is considered learnable if an algorithm exists that, with arbitrarily high probability, can approximate the ground truth function arbitrarily well, when provided a sufficient amount of training data.

A cornerstone of machine learning theory is the well-established no-free-lunch (NFL) theorem for supervised learning (Wolpert, 1996), which states that the hypothesis space of all mathematical functions that map from data to labels is not PAC-learnable. As a consequence, any learning algorithm must inevitably possess an inductive bias. In the light of the NFL theorem, Sterkenburg and Grünwald (2021) have raised the question of how to justify inductive bias.

While these fundamentals of supervised machine learning have not been fully formally transferred to anomaly detection, Siddiqui et al. (2016) have taken a step in this direction and formulated a version of PAC-learnability for unsupervised outlier detection, which yet does not cover semi-supervised anomaly detection as pursued here. In Supplement I, we formally phrase a version of PAC learnability for anomaly detection, along with a conjectured, yet unproven, NFL theorem. Assuming this conjectured NFL theorem, the inductive bias of an anomaly detection approach implicitly contextualizes normality. For anomaly detection approaches without an explicitly contextualized normality, their inductive bias constitutes an implicit sense of normality, and contextualizing normality is one way to make anomaly detection more transparent.

In the *SpyCAT* approach proposed here, we justify inductive bias by the domain knowledge manifested in assumptions (A1)–(A4). This approach is also in line with

recent work in supervised medical image analysis that adjusts inductive bias towards making deep learning models or their interpretable output compliant with existing domain knowledge (Maier et al., 2019; Mahapatra et al., 2022; Wang et al., 2024a; Schuhmacher et al., 2022).

4. Method

The *SpyCAT* approach translates assumptions (A1)–(A4) from Section 3 into an anomaly detection approach. Specifically, our method works by decomposing the complete 3D volume image $\mathcal{I}$ of a brain MRI into cubic patches. For a voxel position p within an image $\mathcal{I}$, we denote the patch originating at p by $\mathbf{I}_p \in \mathbb{R}^{q \times q \times q}$. Throughout this manuscript, we work with a patch size of $q = 16$ unless stated otherwise, which matches assumption (A1). Our approach builds on inferring an anomaly map for individual patches, so that an anomaly map $\mathbf{A}_{\text{map}}$ for a complete MRI image $\mathcal{I}$ can be assembled from a patch-decomposition.

In general, the basic idea is that a vector-quantized variational autoencoder (VQ-VAE) (Oord et al., 2017), which consists of an encoder $V_{\text{enc}}: \mathbb{R}^{q \times q \times q} \rightarrow \mathbf{E}$ and a corresponding decoder V_{dec} , is trained to reliably compress and reconstruct either lesional or non-lesional input images $\mathbf{I}_p$ through a suitable embedding space $\mathbf{E}$. Correspondingly, we expect to obtain an accurate reconstruction of $\mathbf{I}_p$ by computing $\mathbf{I}'_p := V_{\text{dec}}(V_{\text{enc}}(\mathbf{I}_p))$. Following assumption (A2), we can set $\mathbf{I}'_p$ in contrast with a *counterfactual* version of the patch at position p , denoted by $\mathbf{C}_p$, that is to be expected from neighboring brain regions, according to (A3). To obtain this counterfactual $\mathbf{C}_p$, we denote n as one patch position that is adjacent to position p , so that $\mathbf{I}_n$ denotes an image cube adjacent (specifically, *above* p as indicated in Figure 1) to $\mathbf{I}_p$. We then train a transformer network T that utilizes the encoded embedding of $\mathbf{I}_n$ as context information to predict the counterfactual embedding, so that

$$\mathbf{C}_p := V_{\text{dec}}(T(V_{\text{enc}}(\mathbf{I}_n)))$$

yields the desired counterfactual image that is expected from neighboring brain regions, as postulated in (A2). The difference image $|\mathbf{I}'_p - \mathbf{C}_p|$ then yields the anomaly map at position p : If the patch at position p looks as expected from neighboring position n in a non-lesional brain MRI, then $\mathbf{I}'_p$ and $\mathbf{C}_p$ will coincide, yielding an all-zero anomaly map. If, however, a lesion is present at position p , the lesion will be missing in $\mathbf{C}_p$, thus uncovering the lesion in the difference image of the anomaly map.

As stated so far, the transformer T does not account for the brain region in which the embedding of the counterfactual image is to be generated. When dealing with brain MRIs registered into the coordinate system of a reference brain, the location information p is highly informative towards the morphology that is to be reconstructed, as

postulated in assumption (A4). In a transformer network, this can be realized easily: We can simply pass the spatial patch coordinates $p = (x_p, y_p, z_p)$ as additional context information to the transformer, so that we obtain an improved counterfactual image from

$$\mathbf{C}_p := V_{\text{dec}}(T_{\text{perf}}(x_p, y_p, z_p, V_{\text{enc}}(\mathbf{I}_n))).$$

We additionally replaced the transformer architecture by an autoregressive performer (Choromanski et al., 2020), and denote this position-enhanced performer as T_{perf} .

Overall, the training process for our complete anomaly detection network involves two steps. First, the VQ-VAE is trained to bring images into a compressed representation using an encoder, and also to obtain a decoder for image reconstruction. In the second step, a transformer learns the prediction of lesion-free patches at a certain position in the image, and corrects the generated embedding space accordingly.

Finally, the embedding space $\mathbf{E}$ of the variational autoencoder requires some consideration. While the transformer (or performer) network could in principle operate on the level of raw or naively discretized image patches rather than in embedding space, the compression by a VAE has the important role to make the transformer network T_{perf} more data-efficient. As discussed in the next paragraph, this requires an appropriate balance between a sufficient level of compression and a high reconstruction quality.

Obtaining compressed patch representations The key idea behind employing the VQ-VAE (Oord et al., 2017) is to learn a compressed, dictionary-like representation of patches in an embedding space $\mathbf{E}$ that facilitates accurate reconstruction of patches. At the same time, this embedding provides favorable context for a transformer network to predict the embedding of the neighboring patch. To this end, our main building blocks are $K = 8192$ codebook vectors $e_0, \dots, e_{K-1} \subset \mathbb{R}^D$ with a feature depth of $D = 256$ to be inferred from training a VQ-VAE. The original VQ-VAE (Oord et al., 2017) represents the complete input image through a $\mathbb{R}^{K \times D}$ feature matrix, and maps an input to a single D -dimensional codebook vector through a one-hot-encoding of the rows of the feature matrix. In our setting, representing one input patch by a single codebook vector would be too coarse grained for the dedicated purpose of the representation, namely to predict and fully reconstruct a complete neighboring patch. As a more fine-grained and more informative representation, we train the encoder V_{enc} to yield a representation comprising $4 \times 4 \times 4$ codebook vectors, along with a corresponding decoder V_{dec} , which learns to reconstruct an image $\mathbf{I}'_p$ from the embedding space.

If we consider the codebook vectors $e_0, \dots, e_{K-1}$ as fixed after training, we can now formally identify our com-

pressed patch embedding space as

$$\mathbf{E} = \mathbb{I}^{4 \times 4 \times 4},$$

where $\mathbb{I} = \{0, 1, \dots, 8191\}$ denotes the index space of the codebook vectors. In particular, this embedding space allows us to efficiently represent a single patch by 64 integer values that facilitate accurate patch reconstruction. At the same time, the representation based on codebook indices resembles token representations of large language models (Devlin et al., 2018; Touvron et al., 2023) and thus promises to work favorably as an input encoding for transformer networks.

Generating pseudo-lesional patches For our approach to work as intended, V_{enc} and V_{dec} must be able to embed and reconstruct lesions in the autoencoder reconstruction $\mathbf{I}'_p$. The counterfactual reconstruction $\mathbf{C}_p$, obtained from the neighboring patch, is by contrast expected to be lesion-free, as postulated in assumption (A2). However, there is only a very small set of lesional patches available in the dataset, which is far from sufficient to learn their proper reconstruction. Furthermore, MRIs are very heterogeneous and they can have many artifacts. For example, in the SWI sequence, certain sections are commonly masked out (Schönlau, 2018). In order for such artifacts to be considered normal, they must also be present in the embedding space. This in turn requires to enhance the reconstruction capabilities beyond non-lesional patches.

We address these problems by augmenting the dataset by generating *pseudo-lesional* patches. In these augmentation patches, we add random signal that shares characteristics of lesions in the form of Perlin noise (Perlin, 1985) to each image patch $\mathbf{I}_p$. This allows us to use only MRIs marked as non-lesional MRIs for training the VQ-VAE and still achieve sufficient reconstructability at the patch level. In VQ-VAE training, image patches $\mathbf{I}_p$ are sampled at randomly chosen positions in the MRI image.

Perlin noise is a type of gradient noise that creates smooth transitions between points in the grid. For each patch, Perlin noise for image $\mathbf{I}_p$ is generated using Algorithm 1 as follows: First, 3D Perlin noise is generated using the implementation of (Vigier, 2020) (see Figure 2a).

All non-negative voxel intensities of **pn** (Figure 2a) are replaced with the image information from $\mathbf{I}_p$, while negative voxel intensities of **pn** are replaced with the minimum patch value of $\mathbf{I}_p$ (see Figure 2c). A 3D box with side lengths of 3 or 5 is cut out at a random position in the resulting patch (from Figure 2c) and inserted at the same position in $\mathbf{I}_p$ (see Figure 2d). This creates small artifacts that resemble lesions at the patch level.

Perlin noise augmentation is applied to all MRI modalities used in this work (SWI, FLAIR, T1, and T2), so that the VQ-VAE learns to reliably reconstruct such pseudo-lesional image patterns across sequences.

Algorithm 1 *Goal:* Generate a pseudo-lesional image patch by replacing a small cubic region of the input patch with Perlin-noise-derived intensities. *Inputs:* **img_patch**: image patch $\mathbf{I}_p \in \mathbb{R}^{16 \times 16 \times 16}$ sampled at a random position in an MRI volume; **sr**: List of Perlin noise window side lengths (in voxels), here $[3, 5]$. *Internal variables:* **pn**: a 3D Perlin noise volume of the same shape as **img_patch**; **pn_c**: a copy of **pn** that keeps the original sign of each voxel after **pn** is overwritten in place; **s**: the randomly chosen side length of the inserted cube; **ss** = $\lfloor \mathbf{s}/2 \rfloor$: its half side length; **(x, y, z)**: the randomly chosen center voxel of the cube inside **img_patch**. *Output:* the modified **img_patch** containing a small cube of Perlin-noise-derived intensities that resembles a lesion at the patch level.

```

1: procedure GENERATEPERLINNOISEATPATCH(img_patch, sr  $\leftarrow [3, 5]$ )
2:   pn  $\leftarrow$  generate_perlin_noise_3d()           ▷ Generate a 3D Perlin noise volume of the same shape as img_patch (Fig. 2a)
3:   pn_c  $\leftarrow$  pn.clone()                         ▷ Keep an unmodified copy of pn to retain the original sign of each voxel
4:   pn[pn_c < 0]  $\leftarrow$  img_patch.min()             ▷ Negative voxels of pn are overwritten with the minimum of img_patch
5:   pn[pn_c  $\geq$  0]  $\leftarrow$  img_patch[pn_c  $\geq$  0]       ▷ Non-negative pn voxels take img_patch's corresponding intensity (Fig. 2c)
6:   s  $\leftarrow$  random.choice(sr)                     ▷ Randomly sample the side length s  $\in$  sr of the cubic cutout
7:   ss  $\leftarrow$  int(np.floor(s/2))                   ▷ Half side length, used to center the cube around (x, y, z)
8:   x  $\leftarrow$  random.randint(ss, 16 - ss - 1)       ▷ Random x-coordinate of the cube center
9:   y  $\leftarrow$  random.randint(ss, 16 - ss - 1)       ▷ Random y-coordinate of the cube center
10:  z  $\leftarrow$  random.randint(ss, 16 - ss - 1)       ▷ Random z-coordinate of the cube center
11:  img_patch[ :, x-ss : x+ss+1, y-ss : y+ss+1, z-ss : z+ss+1]  $\leftarrow$  pn[ :, x-ss : x+ss+1, y-ss : y+ss+1, z-ss : z+ss+1]
    ▷ Copy the cubic region of pn centered at (x, y, z) into img_patch at the same coordinates (Fig. 2d)
12:  return img_patch                               ▷ Augmented (pseudo-lesional) image patch

```

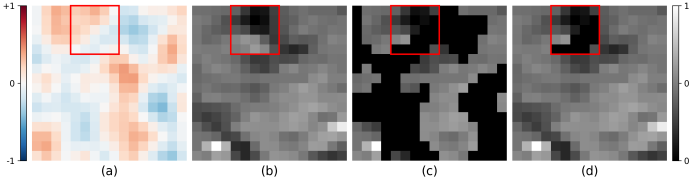


Figure 2: From left to right: (a) Perlin noise **pn**; (b) Image patch $\mathbf{I}_p$ with coronal view; (c) All non-negative voxel intensities of **pn** are replaced with the image information from $\mathbf{I}_p$, while negative voxel intensities of **pn** are replaced with the minimum patch value of $\mathbf{I}_p$; (d) A 3D box with side lengths 3 or 5 (randomly chosen) is cut out of the image in (c) (indicated by the red rectangle) and inserted into the original image $\mathbf{I}_p$ at the same position (indicated by the red rectangle in (b)). The colorbar on the left applies to panel (a) and shows the Perlin noise intensity range $[-1, 1]$ using a diverging colormap, so that negative and positive voxels of **pn** are visually distinguishable. The colorbar on the right applies to panels (b), (c), and (d), which are shown in grayscale with min/max normalization to $[0, 1]$ per image.

Predicting counterfactual neighboring patches The purpose of the VQ-VAE is to obtain a compressed patch representation which facilitates the transformer-based prediction of counterfactually non-lesional neighboring patches following assumptions (A2) and (A3). We here employ the FAVOR+ algorithm introduced in the Performer Framework (Choromanski et al., 2020), which evades the quadratic complexity of the dependencies between query, key, and value vectors, and realizes an attention mechanism with linear time complexity. We train the transformer network T_{perf}

on sequences composed of three parts, assuming that the latent representation of patch position p is to be predicted from a neighboring patch position n :

- The index coordinates x_p, y_p, z_p of the patch position p within the registered coordinate system of the underlying MRI image $\mathcal{I}$, which is encoded through a simple vocabulary embedding,
- the flattened latent space representation of the neighboring patch $z_n := V_{\text{enc}}(\mathbf{I}_n)$, and
- the flattened latent space representation of the patch p itself, i.e., $z_p := V_{\text{enc}}(\mathbf{I}_p)$.

In total, each patch is represented by a sequence composed of $131 = 3 + 64 + 64$ tokens. During inference, only the position embedding and the neighborhood patch embedding are given. Thus, the embedding z_p is predicted autoregressively, resulting in the counterfactual embedding.

The data for training T_{perf} are obtained by drawing image patches $\mathbf{I}_p$ at random positions x_p, y_p, z_p from the training dataset along with an image patch $\mathbf{I}_n$ at a neighboring position p . The patch size of $16 \times 16 \times 16$ as well as the latent space dimension of $4 \times 4 \times 4$ and the codebook size of $K = 8192$ were used constantly throughout all experiments.

To analyze a complete MRI image and determine whether a lesion is present or not, the image is decomposed into patches along a $16 \times 16 \times 16$ voxel grid, where each point p on the grid yields an image patch $\mathbf{I}_p$ along with one neighboring patch $\mathbf{I}_n$. Each $\mathbf{I}_p$ and $\mathbf{I}_n$ is passed through V_{enc} to obtain embeddings z_p and z_n , respectively. We then use T_{perf} to obtain a counterfactual embedding z_c .

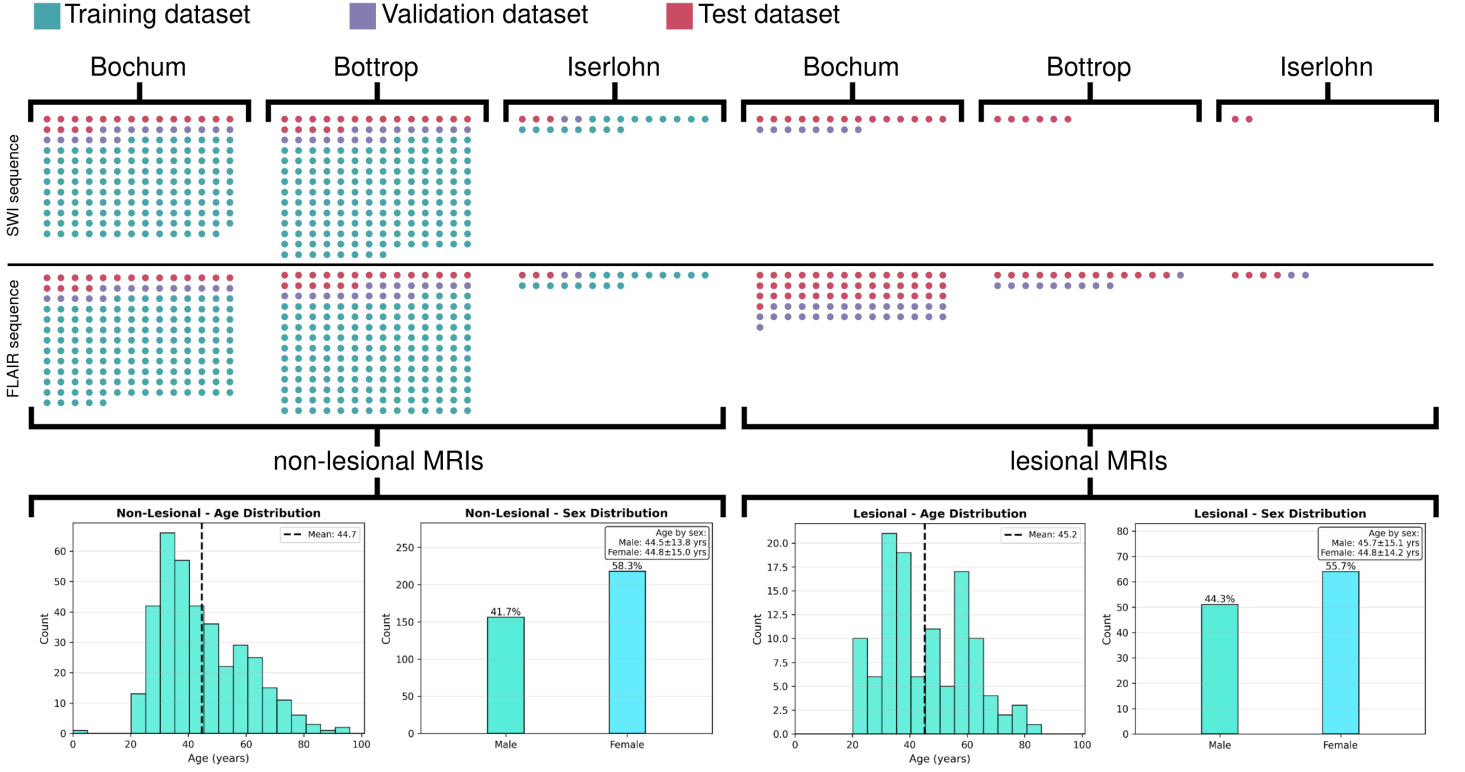


Figure 3: Division of the dataset into training, validation and test data with patient demographics. The MRIs are separated at study level and come from three different hospitals. Age and sex distributions are shown for each subset. SWI sequences are used for cavernoma detection and FLAIR sequences for post-resection cavity detection.

The performer autoregressively replaces all elements given by p and z_n in the sequence that are below a threshold ϕ . The resulting counterfactual embedding space z_c is then passed as input to V_{dec} along with the original embedding space z_p . This produces two reconstructions I'_p and C_p , which are then simply subtracted from each other to obtain a patch anomaly map. All patch anomaly maps can be mapped back to their respective grid positions, which yields the whole-brain anomaly map A_{map} .

5. Experiments

We evaluated *SpyCAT* on three dataset pairs: First, an in-house dataset of brain MRIs related to two types of epileptogenic lesions, henceforth referred to as the *AIM.E dataset*. Second, a dataset obtained by merging the public IXI dataset with the BraTS 2021 dataset for primary brain tumors (gliomas), henceforth referred to as the *IXI-BraTS dataset*. Third, a dataset obtained by merging the IXI dataset with the BraTS-METS 2023 dataset for brain metastases, henceforth referred to as the *IXI-BraTS-METS dataset*. For each dataset and sequence, we train one dedicated model: For the AIM.E dataset, we train separate models for FLAIR sequences (post-resection cavities) and SWI sequences (cavernomas). For the IXI-BraTS and IXI-BraTS-METS datasets, we train models on T2 sequences.

Only non-lesional studies are used for training. More specifically, across all experiments conducted in this manuscript, initial hyperparameter optimization was performed on the validation dataset, while all results including ablation studies are reported based on the test dataset.

5.1 Datasets

AIM.E dataset The AIM.E dataset consists of MRIs that were collected in three hospitals (Bochum, Bottrop, Iserlohn) on Siemens 3 Tesla scanners over a period of 11 years in the time period between 2010 and 2021. Each study contains at least one of the following sequences: 3D-T1 ($1 \times 1 \times 1$ mm), coronal and axial T2 (2 mm), 3D-FLAIR ($1 \times 1 \times 1$ mm) and axial SWI (2 mm) sequence (Wellmer et al., 2013). All MRI studies were evaluated by a physician by first categorizing them into non-lesional vs. lesional. Subsequently, lesions observed in the lesional MRIs were annotated with a singular reference point and given a class label identifying the type of lesion. Within the validation and test datasets, we only include studies where precisely one unambiguously labeled lesion is identified.

For training, only MRIs that have not been labeled *cavernoma* or *post-resection cavity* are used. We consider these as *normal*. For validation and testing, we use lesional and *normal* MRIs. The exact division between training, val-

validation and test dataset is shown in Figure 3. We consider the SWI sequence for cavernoma detection and the FLAIR sequence for post-resection cavity detection (Wellmer et al., 2013; Wang et al., 2017). The division into training, validation, and test dataset takes place at study level, so that there are different numbers for FLAIR and SWI sequences. The dataset comprises 304 non-lesional studies for training, with 35 non-lesional studies for validation and 40 for testing. For cavernoma evaluation on SWI sequences, 8 lesional studies were used for validation and 22 for testing. For post-resection cavity evaluation on FLAIR sequences, 41 lesional studies were allocated to validation and 60 to testing. Demographic characteristics including age at scan time and sex distribution for each data split are provided in Figure 3. Supplement IV shows relevant 2D sections of the 22 validation MRIs for the cavernoma dataset.

IXI-BraTS-Dataset We used the IXI dataset² as a resource for obtaining 581 non-lesional MRI studies, each containing a T1 and a T2 sequence. As a second resource, we utilized the BraTS 2021 dataset (Menze et al., 2015; Bakas et al., 2017, 2018), where each study consists of a T1, T2, and FLAIR sequence accompanied by a segmentation map of five regions (no lesion, non-enhancing tumor core, peritumoral edema, and GD-enhancing tumor). All segmentation values greater than 0 are considered lesional, which leads to binarization (0: non-lesional; 1: lesional). From the BraTS 2021 dataset (1251 total studies), we randomly sampled 581 lesional studies before any model development. These were split into 20% training (116 studies), 40% validation (232 studies), and 40% test (232 studies). For *SpyCAT* and other semi-supervised methods, only the non-lesional IXI data were used for training the VQ-VAE and performer network, while the BraTS lesional studies served exclusively for evaluation. The training split of BraTS data was reserved for potential supervised baseline comparisons but was not used in this study. The IXI non-lesional data were split into 60% training (348 studies), 20% validation (116 studies), and 20% test (117 studies).

IXI-BraTS-METS Dataset We used the same IXI dataset as described above for obtaining non-lesional MRI studies. As a lesional resource, we utilized the BraTS-METS 2023 dataset (Moawad et al., 2023, 2024), which focuses on brain metastases segmentation. Each study consists of T1, T2, FLAIR, and T1 post-contrast sequences accompanied by segmentation masks annotating metastatic lesions. All segmentation values greater than 0 are considered lesional, which leads to binarization (0: non-lesional; 1: lesional). From the BraTS-METS 2023 dataset, we used 328 lesional studies, which were split into validation (96 studies) and test (232 studies). For *SpyCAT* and other semi-supervised meth-

ods, only the non-lesional IXI data was used for training the VQ-VAE and performer network, while the BraTS-METS lesional studies served exclusively for evaluation. The same IXI non-lesional data splits described in the IXI-BraTS section were used.

5.2 Preprocessing

The preprocessing pipeline consists of 4 steps: First, MRIs are registered to the MNI ICBM 2009c Nonlinear Symmetric template (Fonov et al., 2009, 2011) with a size of $193 \times 229 \times 193$ using a rigid transformation (versor) implemented in the SimpleITK framework (Beare et al., 2018; Yaniv et al., 2018; Lowekamp et al., 2013). The T1-weighted sequence of each study is registered to the T1-weighted template sequence, and the resulting transformation matrix is saved and subsequently applied to all other sequences within the study (T2, FLAIR, SWI, etc.) to ensure consistent spatial alignment. Second, skull stripping is performed using the MNI ICBM 2009c template mask. Third, we apply bias field correction using (Sled et al., 1998; Tustison et al., 2010). As a final preprocessing step, we perform histogram normalization following (Nyul et al., 2000).

5.3 Implementation details

Autoencoder The topology of the autoencoder follows Oord et al. (2017) and consists of two convolutional layers with *LeakyReLU* activation function in the encoder V_{enc} and two transposed convolutional layers using *LeakyReLU* as activation function in the decoder V_{dec} , each with dimensions 128 and 256 and six residual layers. We used the Adam optimizer (Kingma and Ba, 2014) for training with a learning rate of $\ell_V = 10^{-5}$, a batch size of 256 and an exponential weight decay of $\gamma_V = 0.95$. Training is performed over 50 epochs, extracting 10,000,000 image patches during each epoch.

Transformer The performer architecture in this paper is based on (Choromanski et al., 2020) and uses six layers, each with eight attention heads and 512 dimensions. The Adam optimizer (Kingma and Ba, 2014) is used with a learning rate of $\ell_T = 10^{-5}$. The training is carried out over 50 epochs, during each of which 1,000,000 image patches $\mathbf{I}_p$ and neighborhood patches $\mathbf{I}_n$ are extracted. The batch size is 256. During the testing phase, the weights from the performer epoch that correspond to the highest validation AUROC metric were utilized.

5.4 Validation Approach

Validation involves the two main constituents of (i) identifying appropriate validation metrics and (ii) identifying appropriate reference methods.

2. <https://brain-development.org/ixi-dataset/>

Pair	Seq.	Model	Classification (<i>ImLC</i>) [%]					Detection (<i>ObD</i>)		
			ACC	AUROC	AP	SENS	SPEC	TP	FN	TP + FN
IXI-BRATS-METS	T2	f-AnoGAN	74.69 _[74.8,79.3]	81.50 _[81.6,84.6]	85.94 _[86.0,88.1]	65.08 _[65.5,73.3]	87.42 _[82.5,87.4]	97	135	232
		β -VAE	68.55 _[67.9,70.0]	75.36 _[75.1,76.9]	81.83 _[81.9,83.1]	65.94 _[52.8,67.2]	72.00 _[72.1,87.8]	113	119	232
		DenosingAE	57.27 _[55.3,60.1]	34.65 _[34.7,37.9]	47.73 _[46.3,47.8]	99.14 _[96.3,99.5]	1.71 _[1.2,5.5]	81	151	232
		PatchCore	71.74 _[71.9,74.2]	74.57 _[74.2,74.6]	72.49 _[72.2,72.9]	75.86 _[76.8,94.4]	66.66 _[47.4,65.3]	141	91	232
		SpyCAT (Ours)	77.39 _[73.2,77.9]	85.22 _[81.5,85.2]	89.78 _[86.9,89.8]	72.41 _[65.0,73.2]	83.99 _[78.3,84.9]	168	64	232
IXI-BraTS	T2	f-AnoGAN	71.25 _[70.8,75.5]	76.55 _[75.6,82.6]	84.18 _[83.4,87.7]	61.21 _[61.2,66.9]	84.57 _[83.5,86.7]	152	80	232
		β -VAE	80.83 _[78.7,82.2]	86.00 _[85.0,87.4]	91.11 _[90.7,92.0]	72.48 _[68.9,75.6]	91.14 _[86.1,92.0]	197	35	232
		DenosingAE	55.77 _[55.1,57.9]	45.30 _[44.5,47.0]	53.69 _[52.6,55.5]	83.18 _[74.6,89.7]	19.42 _[15.6,29.2]	131	101	232
		PatchCore	70.27 _[70.4,74.3]	75.88 _[75.7,78.2]	77.31 _[76.6,79.6]	68.10 _[68.3,93.3]	73.14 _[49.3,73.1]	191	41	232
		SpyCAT (Ours)	78.38 _[78.8,89.5]	85.74 _[86.0,94.4]	91.14 _[91.3,96.5]	68.53 _[69.2,86.8]	91.43 _[91.5,93.1]	210	22	232
AIM.E-PRC	FLAIR	f-AnoGAN	70.59 _[66.8,70.5]	75.79 _[72.8,75.7]	85.36 _[83.3,85.3]	51.67 _[46.9,51.7]	97.61 _[95.2,97.5]	25	35	60
		β -VAE	69.60 _[64.9,69.6]	75.91 _[73.1,75.8]	83.36 _[79.5,83.2]	58.33 _[47.3,58.3]	85.71 _[83.5,90.2]	37	23	60
		DenosingAE	68.62 _[66.8,76.1]	68.53 _[65.6,75.7]	75.89 _[75.2,81.5]	71.66 _[68.5,76.4]	64.28 _[64.3,75.6]	33	27	60
		PatchCore	74.51 _[66.8,74.2]	79.68 _[71.8,79.4]	83.82 _[78.5,83.6]	70.00 _[53.3,71.2]	73.14 _[77.2,81.2]	43	17	60
		SpyCAT (Ours)	75.49 _[75.6,80.3]	78.73 _[78.9,84.1]	86.43 _[84.3,86.7]	71.67 _[72.2,86.4]	80.95 _[73.9,80.7]	44	16	60
AIM.E-Cav.	SWI	f-AnoGAN	66.12 _[67.0,74.1]	74.50 _[75.4,82.4]	67.49 _[67.6,70.7]	95.45 _[90.9,95.2]	0.50 _[50.7,65.0]	18	4	22
		β -VAE	70.97 _[71.1,77.3]	83.00 _[83.5,83.9]	74.66 _[74.7,76.1]	95.45 _[82.2,95.1]	57.50 _[57.9,70.6]	19	3	22
		DenosingAE	74.19 _[72.7,74.2]	62.73 _[62.8,65.5]	56.46 _[49.7,56.3]	50.00 _[50.2,54.5]	87.50 _[82.6,87.4]	16	6	22
		PatchCore	77.42 _[77.6,82.2]	73.63 _[73.9,83.2]	59.30 _[59.6,72.7]	68.18 _[68.6,81.6]	82.49 _[80.1,84.9]	18	4	22
		SpyCAT (Ours)	83.87 _[79.2,83.7]	83.30 _[79.0,83.2]	79.39 _[75.8,79.3]	68.18 _[68.4,72.7]	92.50 _[82.7,92.3]	21	1	22

Table 1: Comparison of our approach with f-AnoGAN, β -VAE, PatchCore, and DenosingAE on the three pairs of datasets (IXI-BraTS-METS, IXI-BraTS and AIM.E dataset). Values are reported as metric_[95% CI lower,upper] computed via bootstrap (1,000 resamples). On the IXI-BraTS-METS and IXI-BraTS dataset, MRIs of the T2 sequence are used. In the AIM.E dataset, the cavernoma evaluation (here abbreviated as *Cav.*) is performed on the SWI data and the post-resection cavity evaluation (here abbreviated as *PRC*) is shown on the FLAIR sequence. For each dataset and metric, the best-performing result is shown in bold.

Validation metrics We used the *metrics reloaded* framework (Maier-Hein et al., 2024) as our basis for identifying appropriate validation measures. Since *SpyCAT* is an anomaly detection approach, and due to the nature of epileptogenic lesions as well as the point-of-interest annotations available for the AIM.E dataset, none of the use cases from (Maier-Hein et al., 2024) are immediately applicable in our setting. Following the framework, we identified two relevant validation categories:

- Image-level Classification (*ImLC*):** We measure *ImLC* performance of distinguishing non-lesional MRIs from lesional ones. We identified the Area Under Receiver Operating Curve (AUROC) as a suitable multi-threshold metric, using the mean across all voxels of the anomaly map $\mathbf{A}_{\text{map}}$ as a natural threshold variable, which we denote as μ_{map} . Analogously to AUROC, we also report average precision (AP). Complementary to these two multi-threshold metrics, we also report accuracy (ACC), sensitivity (SENS) and specificity (SPEC) as single-threshold metrics. To obtain a single threshold, we determine the optimal threshold for μ_{map} by maximizing Youden’s J statistic, i.e., the difference between true positive rate and false positive rate, across the validation dataset. This threshold is then applied as a constant to the complete test dataset to calculate ACC (image-level accuracy), SENS (image-level sensitivity), and SPEC (image-level specificity).
- Object Detection (*ObD*):** Following the process from

(Maier-Hein et al., 2024), we first define a *localization criterion*. We use different localization criteria for the IXI-BraTS and IXI-BraTS-METS datasets versus the AIM.E dataset due to the very different nature of the lesions and the given annotations.

Within our validation and test sets, each MRI contains at most one lesion, ensuring that detection is evaluated at the per-image level with unambiguous ground truth. This constraint was enforced by excluding studies with multiple or ambiguously annotated lesions from evaluation, and allows us to measure detection performance as true positives (TP) when any of the k highest intensity voxels of $\mathbf{A}_{\text{map}}$ satisfies the localization criterion, and false negatives (FN) otherwise.

For the BraTS and BraTS-METS dataset, we note that well-established measures exist for validating supervised segmentation methods, which however are not applicable to the anomaly detection approaches investigated here. As a localization criterion for the BraTS and BraTS-METS dataset, we considered a tumor lesion as detected if any of the k highest intensity voxels of $\mathbf{A}_{\text{map}}$ is located within the ground truth segmentation. Different small values of k were tested through ablation, see Section 6.1 below.

Since the AIM.E dataset lacks ground truth segmentations and possesses only point-of-interest annotations, an additional detection radius parameter needs to be introduced as a localization criterion. Here, we consider

a lesion to be detected if one of the k highest intensity voxels of $\mathbf{A}_{\text{map}}$ is located within a fixed radius of r . As for k , appropriate values for r are subject to ablation, see Section 6.1, where we varied r between $r = 3$ mm and $r = 9$ mm. This detection threshold radius is a much more conservative choice than in the related study by Spitzer et al. (2022), which tolerated a radius of $r = 20$ mm for the detection of focal cortical dysplasia lesions.

We note that neither true negatives nor false positives can occur when evaluating *ObD*.

Statistical Analysis All performance metrics are reported with 95% bootstrap confidence intervals (1,000 resamples using the percentile method). For Table 1, confidence intervals were computed on the independent test set. For Table 2 (ablation studies), confidence intervals were computed on the test dataset of the cavernoma task.

Reference methods We employed four semi-supervised approaches as reference methods for comparing our approach to the state-of-the-art: a 3D f-AnoGAN (Simarro et al., 2020), a β -VAE (Loizillon et al., 2024), PatchCore (Roth et al., 2021), and DenoisingAE (Kascenas et al., 2023). The first two methods operate at the 3D volume level, processing entire MRI volumes without slice or patch extraction. In contrast, DenoisingAE and PatchCore work at the 2D slice level, where each 3D volume is processed as a collection of individual slices. All methods are trained on non-lesional MRIs and tested on lesional MRIs (semi-supervised approach). f-AnoGAN, β -VAE, and DenoisingAE share the same underlying idea as our approach, namely that the lesional areas of the MRI, unlike *normal* healthy areas, cannot be reconstructed by the network, so an anomaly map is obtained by comparing the input image with the output reconstruction. PatchCore, in contrast, uses a memory bank approach to detect anomalies based on feature distance to normal training samples.

The implementation of the 3D f-AnoGAN utilized here is based on (Simarro et al., 2020), but does not reduce the original images to $64 \times 64 \times 64$, but to $128 \times 128 \times 128$, in order to have a chance to detect small lesions at all. In the implementation, a batch size of 2 is used. With the β -VAE, the full image resolution is used as input in contrast to the 3D f-AnoGAN. All other parameters follow (Simarro et al., 2020) and (Loizillon et al., 2024).

For PatchCore and DenoisingAE, we adapted the methods to our 3D MRI data by extracting coronal slices from each 3D volume, resulting in 2D images of size $1 \times 193 \times 193$. Each slice was processed independently through the respective method, and the resulting 2D anomaly maps were stacked together to reconstruct the full 3D anomaly volume. This 2D slice-based approach was necessary as both

methods were originally designed for 2D image data. All other parameters follow the original implementations, using code from the official GitHub repositories.

6. Results

The performance of *SpyCAT* and the reference methods in terms of the *Metrics Reloaded* validation categories of *image classification* (ImLC) and *object detection* (ObD) are shown in Table 1.

Image-Level Classification Performance *SpyCAT* consistently achieves high classification performance across all four evaluation pairs, demonstrating robust generalization across different pathologies and imaging modalities. On the BraTS-METS dataset, *SpyCAT* attains an AUROC of 85.22%, substantially outperforming f-AnoGAN (81.50%), β -VAE (75.36%), PatchCore (74.57%), and DenoisingAE (34.65%). Similarly, on the BraTS glioma dataset, *SpyCAT* reaches 85.74% AUROC, closely matching the performance of β -VAE (86.00%) while maintaining higher specificity (91.43% vs. 91.14%). The average precision (AP) metric further underscores *SpyCAT*'s discriminative capabilities, with values of 89.78% (BraTS-METS) and 91.14% (BraTS) exceeding all baseline methods.

The performance gap between *SpyCAT* and reconstruction-based baselines is particularly pronounced for the epileptogenic lesion datasets. On post-resection cavities (AIM.E-PRC), *SpyCAT* achieves 86.43% AP compared to 85.36% for f-AnoGAN and 83.82% for PatchCore, while maintaining balanced sensitivity (71.67%) and specificity (80.95%). For cavernomas (AIM.E-Cav.), *SpyCAT* demonstrates the largest performance advantage with 83.87% accuracy and 83.30% AUROC, substantially outperforming PatchCore (77.42% accuracy) and f-AnoGAN (66.12% accuracy). The DenoisingAE baseline exhibits near-random performance on BraTS-METS (AUROC 34.65%) and cavernomas (AUROC 62.73%), reflecting its inability to capture complex spatial dependencies without explicit context modeling.

The sensitivity-specificity trade-off reveals important methodological differences between approaches. DenoisingAE consistently achieves extremely high sensitivity (99.14% on BraTS-METS, 83.18% on BraTS) at the cost of severely degraded specificity (1.71% and 19.42% respectively), suggesting a tendency to classify most images as anomalous. In contrast, *SpyCAT* maintains more balanced operating points, with sensitivity ranging from 68.18% to 72.41% and specificity from 80.95% to 92.50% across datasets. This balance is clinically preferable for screening applications, where excessive false positives would overwhelm radiological workflows. The β -VAE baseline demonstrates competitive performance on gliomas (86.00% AUROC) but degrades substantially on metastases (75.36% AUROC), indicating

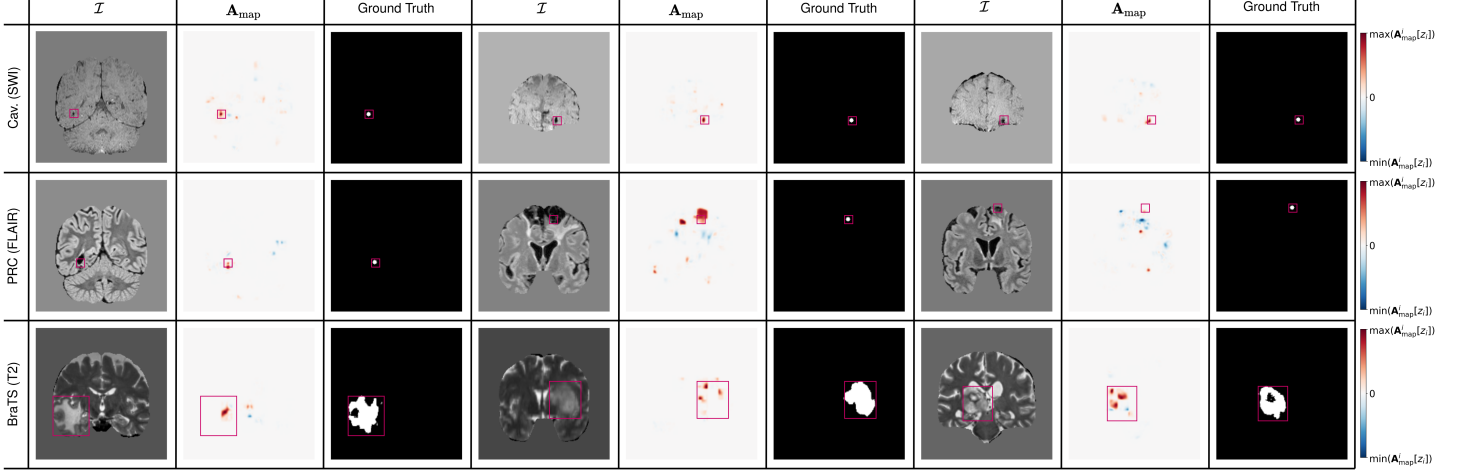


Figure 4: Representation of the output of *SpyCAT* on the three lesion types. Each row corresponds to one lesion entity on its associated MRI sequence: cavernomas on SWI (first row), post-resection cavities on FLAIR (second row), and BraTS tumors on T2 (third row). Ground truth point-of-interest annotations exist for the AIM.E data (cavernomas and post-resection cavities); segmentation maps exist for the BraTS dataset. For each input image $\mathcal{I}$, the corresponding anomaly map $\mathbf{A}_{\text{map}}$, composed of the individual patch predictions, is shown as a coronal slice. Anomaly maps are displayed with a diverging colormap centered on 0, with negative values in blue and positive values in red. To make lesions of widely different magnitudes equally visible across panels, each anomaly map is normalized symmetrically to its own range $[-v_{\text{max}}^i, +v_{\text{max}}^i]$, where v_{max}^i is the largest absolute value of $\mathbf{A}_{\text{map}}$ in the coronal slice z_i of subject i . To better illustrate the location of the ground truth annotations in $\mathcal{I}$ and $\mathbf{A}_{\text{map}}$, bounding boxes were drawn at the same coordinates in both the input image and the corresponding anomaly map.

limited generalization across tumor types with different spatial characteristics.

Object Detection Performance The object detection metrics evaluate whether lesions are successfully localized within images, providing complementary evidence of clinical utility beyond image-level classification. *SpyCAT* achieved the highest detection rate on every dataset; on BraTS gliomas, the closest competing method was β -VAE, with 197/232 (84.9%) detections compared with 210/232 (90.5%) for *SpyCAT*.

The relationship between classification and detection performance varies across methods, highlighting their different strengths. While β -VAE achieves strong AUROC on BraTS gliomas (86.00%), its detection rate (84.9%) is lower than *SpyCAT* (90.5%). This suggests that β -VAE can effectively rank images by abnormality but struggles to precisely locate lesions within the MRI image.

The detection performance on epileptogenic lesions is particularly notable given their subtle appearance. For cavernomas, *SpyCAT* successfully localizes 21/22 lesions (95.5%), compared to 18/22 for both f-AnoGAN and PatchCore, despite cavernomas representing the smallest pathological entities in our evaluation. This high detection rate on subtle lesions validates the hypothesis that spatially context-aware transformers can effectively leverage surrounding anatomical context to amplify weak anomaly signals.

For post-resection cavities, *SpyCAT* detects 44/60 lesions (73.3%), marginally outperforming PatchCore (43/60, 71.7%) but with substantially better classification metrics. This suggests that while both methods can localize cavities when present, *SpyCAT*'s anomaly scores are more discriminative for distinguishing lesional from non-lesional images. The relatively lower detection rate on this lesion type compared to cavernomas (73.3% vs. 95.5%) may reflect the greater morphological heterogeneity of post-surgical changes. Their appearance, in fact, ranges from well-defined fluid-filled cavities to irregular scar tissue with ambiguous boundaries.

A major question that is central to the approach behind *SpyCAT* is how lesion size relates to detection performance. To assess this, we have performed a volume-stratified evaluation of lesion detection performance. Specifically, as detailed in Supplement II, we have compared the performance of *SpyCAT* with the β -VAE for lesions binned into six different classes according to their size. This stratified analysis unveils that *SpyCAT*, compared to β -VAE, strongly improves the detection of small lesions in the BraTS-METS dataset. The detection of larger lesions is improved to a much smaller degree. A similar trend, although somewhat overshadowed by training data limitations, is observed for multiple sclerosis lesions in the MSLesSeg dataset (Guarnera et al., 2025), as detailed in Supplement III.

Qualitative Assessment Figure 4 shows the three classes to be examined, namely *cavernoma*, *post-resection cavity* as well as *tumor* in the BraTS data. For each entity, three images are shown: First, an example image $\mathcal{I}$ (coronal view), second, the corresponding anomaly map $\mathbf{A}_{\text{map}}$, and third, a ground truth map. In the AIM.E dataset, the annotated point-of-interest is shown as ground truth, while for the BraTS dataset the ground truth segmentation is shown.

The displayed results indicate that even the smallest lesions (cavernoma and post-resection cavity) are highlighted well in the anomaly map $\mathbf{A}_{\text{map}}$. It becomes particularly clear that a deviation is visible in contrast to the local neighborhood patch $\mathbf{I}_n$ and the position embedding x_p, y_p, z_p .

Larger lesions such as those in the BraTS tumor dataset significantly exceed the patch size and thus do not match the modeling assumptions underlying *SpyCAT*. Specifically, some neighboring patches $\mathbf{I}_n$ of tumor patches will also contain tumor tissue. Despite this limitation, tumors are also detected in the BraTS dataset. However, as Figure 4 indicates, lesions are only detected at their boundaries, where the tumor edge creates enough contrast between neighboring patches for *SpyCAT* to pick up the anomaly. *SpyCAT*'s spatial context reasoning therefore remains robust even when its underlying assumptions are only partially satisfied. Optimal performance, however, is only achieved when lesions satisfy the spatial locality constraints encoded in our architecture.

6.1 Ablation Studies

All ablation studies were conducted on the test dataset of the cavernoma task using the SWI sequence only, as the validation set was too small for reliable ablation analysis. Note that although Perlin noise augmentation is used for all modalities in our method, the quantitative ablation of Perlin noise was performed on SWI data.

Detection threshold ϕ It is particularly important to determine when the performer considers an element in the sequence to be abnormal, given the position embedding x_p, y_p, z_p and the neighborhood sequence $\mathbf{z}_n$. Since there is no straightforward choice of ϕ that can be justified mathematically or by domain knowledge, we treat ϕ as a parameter. The parameter ϕ determines which elements in the sequence are replaced and which are not. If the value of the element falls below ϕ , then the element is replaced by the element suggested by the performer. In this ablation study, we tested three values $\phi \in \{0.01, 0.001, 0.0001\}$. The results shown in Table 2 indicate that ϕ is relatively robust across the three orders of magnitude of variation. Finetuning ϕ towards a particular setting should be regarded as recommended practice.

Validation radius r Another important parameter in our validation procedure is when a lesion is recognized as a lesion, i.e., which radius r is chosen. In the lack of a natural choice for r , we evaluated the influence of different radii $r \in \{3 \text{ mm}, 5 \text{ mm}, 7 \text{ mm}, 9 \text{ mm}\}$. Reducing the radius strongly towards $r = 3$ leads to detecting only 15 rather than 21 out of the 22 lesions. It should be noted that reducing the detection radius towards or below the size of the lesions reaches a limitation of point-based annotations. Our results can also be viewed in the context of the procedure by (Spitzer et al., 2023), where a more generous radius of $r = 20 \text{ mm}$ was chosen for the detection of focal cortical dysplasia lesions. As a conclusion, r needs to be chosen with the lesion size in mind when validating against point-of-interest annotations.

Validation top k criterion The choice of the parameter k , which describes up to which number of the top k values of the anomaly map $\mathbf{A}_{\text{map}}$ a lesion is considered to be detected, is another validation parameter that does not follow a natural choice. We therefore tested the values $k \in \{3, 5, 10\}$, see Table 2. The detection of 18 lesions at the very conservative choice of $k = 3$, and only three more lesions detected under $k = 10$, underlines the robustness of the anomaly map produced by *SpyCAT*.

Perlin noise In the training process of the VQ-VAE, we use Perlin noise as data augmentation with the intuition that the autoencoder also learns with the data augmentation to restore *lesion-like* tissue and to map it in the embedding space. Comparison with and without the data augmentation can be seen in Table 2, which indicates that *SpyCAT* in general can work without Perlin noise, its introduction does improve performance.

Patch size Since the patch size of the neighborhood patch $\mathbf{I}_n$ and the patch to be predicted $\mathbf{I}_p$ are identical in our model, the size q of the neighborhood patch directly determines the available context information. We systematically evaluated 3D patches with sizes $q \in \{8, 12, 16\}$ regarding their impact on classification and detection performance. The results demonstrate a clear performance degradation with smaller patch sizes: at $q = 12$, accuracy (ACC) decreases by 9.68 %, AUROC decreases by 6.37 %, and average precision (AP) decreases by 8.92 % compared to $q = 16$. The even smaller size $q = 8$ leads to more dramatic losses with ACC= 61.29 % (a decline of 22.58 %), AUROC= 67.72 %, and AP= 53.99 %. These results suggest that sufficiently large spatial context is crucial for the transformer's reliable prediction of normality.

Patch overlap strategy The anomaly map $\mathbf{A}_{\text{map}}$ is obtained by decomposing the MRI volume into non-overlapping $16 \times 16 \times 16$ patches. In this ablation study, we investigated three patch overlap ratios, 0%, 25%, 50%, during the

decomposition process. For voxels covered by multiple overlapping patches, the final anomaly value was computed as the mean of the corresponding patch-wise predictions. As shown in Table 2, introducing 25% overlap yielded modest improvements across all classification metrics: ACC increased by +1.61%, AUROC by +1.33%, AP by +3.22%, sensitivity by +0.91%, and specificity by +1.62%. With 50% overlap, performance gains were more pronounced (ACC +3.22%, AUROC +2.67%, AP +6.44%, sensitivity +1.83%, specificity +3.25%), but at significantly increased computational cost (approximately $4\times$ more patches to process). Notably, detection performance remained identical across all three configurations (21 TP, 1 FN). These results demonstrate that the additional computational overhead yields noticeable, but relatively small performance gains. The neighboring patch I_n in our architecture appears to provide sufficient contextual information, effectively mitigating potential boundary effects without requiring dense overlapping patches, suggesting non-overlapping patches as a reasonable default configuration.

7. Discussion and Conclusion

With *SpyCAT*, we have presented a novel anomaly detection approach that identifies anomalies as deviations from local context, which is realized by combining a transformer-based approach with patch tokenization through a vector-quantized variational autoencoder. We demonstrate that our approach improves the performance of conceptually related state-of-the-art approaches.

The detection of epileptogenic lesions has caught considerable attention in recent years. As reviewed by Walger et al. (2023), major efforts have been spent on the identification of focal cortical dysplasias (FCDs) and their multicentric validation (Kersting et al., 2024). While FCDs have not been included in our present study, the explicit assumptions underlying our anomaly detection apply to FCDs as well, so that it is promising to assess FCD detection using our proposed approach.

In our present work, we use spatial context only in top-down direction. This limitation is mostly for technical reasons, and may be enhanced by incorporating context from different neighboring regions. It is generally conceivable to involve spatial information in a knowledge-guided manner, which is well in line with the modeling based approach behind *SpyCAT*, specifically modeling assumption (A4). This could be realized by utilizing symmetric context if lesions of a certain type are understood to break brain symmetry, e.g. in cortical regions. A further source of knowledge-guided predictive lesion context could be atlas information in MNI space, where one can potentially choose specific context patches in white matter or gray matter regions. Our first attempts to utilize symmetric context in

practice, however, failed to improve lesion detection (data not shown). One likely reason is the insufficient precision of registration to the MNI brain, which may be worthwhile to assess under promising recent progress in brain registration (Wang et al., 2024b).

While *SpyCAT* detects lesions more reliably than reference methods within a suitable detection radius, it is in its current form not able to provide segmentations of the lesions, but only yields localization at the patch level. In principle, this could be addressed by striding the detection patch voxel by voxel; however, this is computationally prohibitively expensive in practice. For our purpose, namely the detection of small lesions, localization at the patch level can be considered sufficient and comes at the further advantage that for validation, only point annotations are required.

A methodological contribution of *SpyCAT* beyond brain MRI image analysis is the use of the VQ-VAE to produce discrete embeddings as encodings of 3D patches. Since it is plausible that transformer networks perform favorably on discrete tokens, this idea appears natural and has been described in similar form for the very different task of modeling of episodic memory in cognitive science (Fayyaz et al., 2022).

Our approach is based on explicit modeling assumptions about what kinds of variances set normal images apart from abnormal ones. In this respect, our approach deviates significantly from most previous anomaly detection methods, which tend to be conceived as generic and application independent approaches. Our specific modeling assumptions (A1–A4) are tailored explicitly towards small lesion detection. As we could show in volume-stratified validation on the BraTS-METS dataset, this translates to a particular strengthening of the detection of small lesions. The less pronounced, but still significant, improvement of large lesion detection that we observe may come as a surprise. Yet, assumptions (A2–A4) are all independent of lesion size, and also improve the detection of large lesions as long as they are sharply differentiated from normal surrounding tissue. Only assumption (A1) inhibits large lesion detection, which we conjecture manifests itself in the anomaly map, where large lesions leave only partial traces, but are still detectable. To assess this, volumetric validation would be required which is beyond the scope of this current study.

The general idea behind our approach is that any anomaly detection algorithm, just like any supervised learning algorithm, inevitably involves if not explicit, then implicit assumptions in the form of an inductive bias. While for supervised learning, this is well-established through the no-free-lunch theorems of supervised learning, these theoretical aspects have not been studied for anomaly detection. In this direction, our work calls for investigating how far and in what form anomaly detection is subject to no-free-lunch

Pair	Seq.	ϕ	r	k	pn	q	ovlp	Classification [%]					Detection		
								ACC	AUROC	AP	SENS	SPEC	TP	FN	TP + FN
AIM.E-Cav.	SWI	$\phi = 0.01$	r = 9	k = 10	y	q = 16	0%	72.58 _[68.0,75.6]	76.02 _[73.6,81.2]	71.25 _[67.6,76.5]	72.73 _[73.0,85.9]	72.50 _[62.9,72.4]	17	5	22
		$\phi = 0.001$	r = 9	k = 10	y	q = 16	0%	83.87 _[79.2,83.7]	83.30 _[79.0,83.2]	79.39 _[75.8,79.3]	68.18 _[68.4,72.7]	92.50 _[82.7,92.3]	21	1	22
		$\phi = 0.0001$	r = 9	k = 10	y	q = 16	0%	79.03 _[79.3,86.9]	79.32 _[79.7,89.8]	78.30 _[78.7,87.2]	68.18 _[68.2,81.1]	85.00 _[85.0,96.9]	16	6	22
		$\phi = 0.001$	r = 3	k = 10	y	q = 16	0%						15	7	22
		$\phi = 0.001$	r = 5	k = 10	y	q = 16	0%						20	2	22
		$\phi = 0.001$	r = 7	k = 10	y	q = 16	0%						20	2	22
		$\phi = 0.001$	r = 9	k = 10	y	q = 16	0%						21	1	22
		$\phi = 0.001$	r = 9	k = 10	y	q = 16	0%						21	1	22
		$\phi = 0.001$	r = 9	k = 10	y	q = 16	0%						19	3	22
		$\phi = 0.001$	r = 9	k = 5	y	q = 16	0%						18	4	22
		$\phi = 0.001$	r = 9	k = 3	y	q = 16	0%								
		$\phi = 0.001$	r = 9	k = 10	y	q = 16	0%	83.87 _[79.2,83.7]	83.30 _[79.0,83.2]	79.39 _[75.8,79.3]	68.18 _[68.4,72.7]	92.50 _[82.7,92.3]	21	1	22
		$\phi = 0.001$	r = 9	k = 10	n	q = 16	0%	79.03 _[71.2,78.8]	76.59 _[72.1,76.5]	74.28 _[68.2,74.1]	59.09 _[57.2,61.0]	90.00 _[86.2,91.9]	19	3	22
		$\phi = 0.001$	r = 9	k = 10	y	q = 16	0%	83.87 _[79.2,83.7]	83.30 _[79.0,83.2]	79.39 _[75.8,79.3]	68.18 _[68.4,72.7]	92.50 _[82.7,92.3]	21	1	22
		$\phi = 0.001$	r = 9	k = 10	y	q = 12	0%	74.19 _[72.7,83.4]	76.93 _[77.1,80.5]	70.47 _[70.7,79.4]	77.27 _[65.0,81.6]	72.50 _[67.5,90.6]	20	2	22
		$\phi = 0.001$	r = 9	k = 10	y	q = 8	0%	61.29 _[59.8,61.5]	67.72 _[65.8,67.8]	53.99 _[52.8,57.7]	90.90 _[90.9,95.2]	45.00 _[40.3,45.0]	17	5	22
		$\phi = 0.001$	r = 9	k = 10	y	q = 16	0%	83.87 _[79.2,83.7]	83.30 _[79.0,83.2]	79.39 _[75.8,79.3]	68.18 _[68.4,72.7]	92.50 _[82.7,92.3]	21	1	22
		$\phi = 0.001$	r = 9	k = 10	y	q = 16	25%	85.48 _[81.6,85.3]	84.63 _[81.2,84.5]	82.61 _[77.7,82.5]	69.09 _[68.3,71.4]	94.12 _[87.7,94.8]	21	1	22
		$\phi = 0.001$	r = 9	k = 10	y	q = 16	50%	87.09 _[84.0,87.0]	85.97 _[83.4,85.7]	85.83 _[79.6,85.7]	70.01 _[68.2,70.1]	95.75 _[92.6,97.4]	21	1	22

Table 2: Results of the ablation studies for the parameters $\phi \in \{0.01, 0.001, 0.0001\}$, $\mathbf{r} \in \{3, 5, 7, 9\}$, $\mathbf{k} \in \{3, 5, 10\}$, query patch size $\mathbf{q} \in \{8, 12, 16\}$, overlap $\mathbf{ovlp} \in \{0\%, 25\%, 50\%\}$, and VQ-VAE training with (y) and without (n) Perlin noise (pn). Note that modifying $\mathbf{r}$ and $\mathbf{k}$ affects detection performance, but not classification metrics, so their ablation in the classification columns remains blank.

theorems.

One practically relevant implication of our modeling-based approach is that the explicit assumptions facilitate explanation of misidentified anomalies. For example, we can explain why only small subregions of the relatively large tumors in the BraTS dataset are identified by *SpyCAT*: major fractions of tumors cover several adjacent patches and thus are inconsistent with our prior assumptions. Similarly, it would in principle be possible to attribute other false negative identifications of patches to further deviations in the neighboring patches, which then fail to properly identify and localize the abnormality. In this sense, the explicit underlying assumptions of our approach warrant a certain degree of explainability, and thus improve the transparency compared to approaches without explicit assumptions.

Finally, another limitation towards clinical practice of our anomaly detection approach is the lack of false positive analysis at the detection level, as potentially excessive false positive detections inflate visual inspection time and thus compromise productivity gains. While false positive analysis is, in principle, available at the image classification level, our present study is limited by the lack of non-lesional images in some of our datasets, most notably the IXI-BraTS and the IXI-BraTS-METS datasets. In the given setting, classification measures may partially represent dataset differences. In future studies, the shortcoming at detection level can potentially be addressed by additional lesion filtering in binarized anomaly maps.

Acknowledgments

This research was funded in part by the humAlne project, funded by the German Ministry of Science and Education

(FKZ 02L19C203) and the FoRUM program of the Medical Faculty of the Ruhr University Bochum (F1089N-2004). We thank Sven Kreienbrock and Tobias Erm for technical assistance.

Ethical Standards

The study was approved by the Ethics Committee of the medical faculty of Ruhr-University Bochum (21-7336).

Conflicts of Interest

We declare that we have no conflicts of interest.

Data availability

The AIM.E dataset is available on request from the authors. *SpyCAT* weights and code can be accessed via <https://doi.org/10.5281/zenodo.16088059> and <https://github.com/johannesSX/SpyCAT> (tag v1.0-melba), respectively.

References

- Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4:170117, September 2017. ISSN 2052-4463. .
- Spyridon Bakas, Mauricio Reyes, Andras Jakab, Ste-

- fan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.
- Christoph Baur, Benedikt Wiestler, Mark Muehlau, Claus Zimmer, Nassir Navab, and Shadi Albarqouni. Modeling healthy anatomy with artificial intelligence for unsupervised anomaly detection in brain MRI. *Radiology: Artificial Intelligence*, 3(3):e190169, 2021.
- Richard Beare, Bradley Lowekamp, and Ziv Yaniv. Image segmentation, registration and characterization in R with SimpleTK. *Journal of Statistical Software*, 86(8):1–35, 2018. . URL <https://www.jstatsoft.org/index.php/jss/article/view/v086i08>.
- Finn Behrendt, Debayan Bhattacharya, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Patched diffusion models for unsupervised anomaly detection in brain MRI. March 2023. .
- Cosmin I. Bercea, Philippe C. Cattin, Julia A. Schnabel, and Julia Wolleb. Denoising diffusion models for anomaly localization in medical images. *Machine Learning for Biomedical Imaging*, 3:687–711, 2025. ISSN 2766-905X. . URL <https://melba-journal.org/2025:030>.
- Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019.
- Simona Bottani, Ninon Burgos, Aurelien Maire, Dario Saracino, Sebastian Ströer, Didier Dormont, and Olivier Colliot. Evaluation of MRI-based machine learning approaches for computer-aided diagnosis of dementia in a clinical data warehouse. *Medical Image Analysis*, 89:102903, 2023. ISSN 1361-8415. . URL <https://www.sciencedirect.com/science/article/pii/S1361841523001639>.
- Christopher Bowles, Chen Qin, Ricardo Guerrero, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. Brain lesion segmentation through image synthesis and outlier detection. *NeuroImage: Clinical*, 16: 643–658, 2017.
- Yu Cai, Weiwen Zhang, Hao Chen, and Kwang-Ting Cheng. MediAnomaly: A comparative study of anomaly detection in medical images. *Medical Image Analysis*, page 103500, 2025.
- Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. *NeuroImage*, 148:77–102, 2017.
- Krzysztof Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers. September 2020. .
- Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: a patch distribution modeling framework for anomaly detection and localization. November 2020. .
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. October 2018. .
- Nima Fathi, Amar Kumar, Brennan Nichyporuk, Mohammad Havaei, and Tal Arbel. Decodex: Confounder detector guidance for improved diffusion-based counterfactual explanations. *arXiv preprint arXiv:2405.09288*, 2024.
- Gian Mario Favero, Parham Saremi, Emily Kaczmarek, Brennan Nichyporuk, and Tal Arbel. Conditional diffusion models are medical image classifiers that provide explainability and uncertainty for free. *arXiv preprint arXiv:2502.03687*, 2025.
- Zahra Fayyaz, Aya Altamimi, Carina Zoellner, Nicole Klein, Oliver T Wolf, Sen Cheng, and Laurenz Wiskott. A model of semantic completion in generative episodic memory. *Neural Computation*, 34(9):1841–1870, 2022.
- Vladimir Fonov, Alan C. Evans, Kelly Botteron, C. Robert Almli, Robert C. McKinstry, and D. Louis Collins. Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1):313–327, 2011. ISSN 1053-8119. . URL <https://www.sciencedirect.com/science/article/pii/S1053811910010062>.
- VS Fonov, AC Evans, RC McKinstry, CR Almli, and DL Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:S102, 2009. ISSN 1053-8119. . URL <https://www.sciencedirect.com/science/article/pii/S1053811909708845>. Organization for Human Brain Mapping 2009 Annual Meeting.
- Ravnoor Singh Gill, Hyo-Min Lee, Benoit Caldaïrou, Seok-Jun Hong, Carmen Barba, Francesco Deleo, Ludovico D’Incerti, Vanessa Cristina Mendes Coelho, Matteo

- Lenge, Mira Semmelroch, Dewi Victoria Schrader, Fabrice Bartolomei, Maxime Guye, Andreas Schulze-Bonhage, Horst Urbach, Kyoo Ho Cho, Fernando Cendes, Renzo Guerrini, Graeme Jackson, R Edward Hogan, Neda Bernasconi, and Andrea Bernasconi. Multicenter validation of a deep learning detection algorithm for focal cortical dysplasia. *Neurology*, 97(16):e1571–e1582, October 2021.
- Francesco Guarnera, Alessia Rondinella, Elena Crispino, Giulia Russo, Clara Di Lorenzo, Davide Maimone, Francesco Pappalardo, and Sabastiano Battiato. MSLesSeg: baseline and benchmarking of a new Multiple Sclerosis Lesion Segmentation dataset. 6 2025. . URL https://springernature.figshare.com/articles/dataset/MSLesSeg_baseline_and_benchmarking_of_a_new_Multiple_Sclerosis_Lesion_Segmentation_dataset/27919209.
- Jia Guo, Shuai Lu, Lize Jia, Weihang Zhang, and Huiqi Li. Encoder-decoder contrast for unsupervised anomaly detection in medical images. *IEEE Transactions on Medical Imaging*, 43(3):1102–1112, 2024. .
- Ahmed Hamada. Br35h: Brain tumor detection 2020, 2020. URL <https://www.kaggle.com/ahmedhamada0/brain-tumor-detection>.
- Hasan Iqbal, Umar Khalid, Jing Hua, and Chen Chen. Unsupervised anomaly detection in medical images using masked diffusion model. May 2023. .
- Antanas Kascenas, Rory Young, Björn Sand Jensen, Nicolas Pugeault, and Alison Q. O’Neil. Anomaly detection via context and local feature matching. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 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. *Medical Image Analysis*, 90:102963, 2023. ISSN 1361-8415. . URL <https://www.sciencedirect.com/science/article/pii/S1361841523002232>.
- Lennart N. Kersting, Lennart Walger, Tobias Bauer, Vadym Gnatkovsky, Fabiane Schuch, Bastian David, Elisabeth Neuhaus, Fee Keil, Anna Tietze, Felix Rosenow, Angela M. Kaindl, Elke Hattingen, Hans-Jürgen Huppertz, Alexander Radbruch, Rainer Surges, and Theodor Rüber. Detection of focal cortical dysplasia: Development and multicentric evaluation of artificial intelligence models. *Epilepsia*, n/a(n/a), 2024. .
- URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/epi.18240>.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. December 2014. .
- Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. CFA: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. June 2022. .
- Sophie Loizillon, Yannick Jacob, Maire Aurélien, Didier Dormont, Olivier Colliot, Ninon Burgos, and APPRIM-AGE Study Group. Detecting brain anomalies in clinical routine with the β -VAE: Feasibility study on age-related white matter hyperintensities. In *Medical Imaging with Deep Learning*, 2024.
- Bradley C Lowekamp, David T Chen, Luis Ibáñez, and Daniel Blezek. The design of SimpleTK. *Frontiers in neuroinformatics*, 7:45, 2013.
- Ruiying Lu, YuJie Wu, Long Tian, Dongsheng Wang, Bo Chen, Xiyang Liu, and Ruimin Hu. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 36:8487–8500, 2023.
- Guoting Luo, Wei Xie, Ronghui Gao, Tao Zheng, Lei Chen, and Huaiqiang Sun. Unsupervised anomaly detection in brain mri: Learning abstract distribution from massive healthy brains. *Computers in Biology and Medicine*, 154:106610, 2023. ISSN 0010-4825. . URL <https://www.sciencedirect.com/science/article/pii/S0010482523000756>.
- Dwarikanath Mahapatra, Alexander Poellinger, and Mauricio Reyes. Interpretability-guided inductive bias for deep learning based medical image. *Medical image analysis*, 81:102551, 2022.
- Andreas K Maier, Christopher Syben, Bernhard Stimpel, Tobias Würfl, Mathis Hoffmann, Frank Schebesch, Weilin Fu, Leonid Mill, Lasse Kling, and Silke Christiansen. Learning with known operators reduces maximum error bounds. *Nature machine intelligence*, 1(8):373–380, 2019.
- Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: recommendations for image analysis validation. *Nature methods*, 21(2):195–212, 2024.
- Jordi Malé, Juan Fortea, Mateus Rozalem Aranha, Yann Heuzé, Neus Martínez-Abadías, and Xavier Sevillano.

- Towards the discovery of down syndrome brain biomarkers using generative models. In *European Conference on Computer Vision*, pages 207–221. Springer, 2024.
- Sergio Naval Marimont and Giacomo Tarroni. Anomaly detection through latent space restoration using vector-quantized variational autoencoders. December 2020. .
- Felix Meissen, Johannes Paetzold, Georgios Kaissis, and Daniel Rueckert. Unsupervised anomaly localization with structural feature-autoencoders. In Spyridon Bakas, Alessandro Crimi, Ujjwal Baid, Sylwia Malec, Monika Pytlarz, Bhakti Baheti, Maximilian Zenk, and Reuben Dorent, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 14–24, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-33842-7.
- Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lenczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Aşatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, Josée Antoniaş Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging*, 34:1993–2024, October 2015. ISSN 1558-254X. .
- Ahmed W. Moawad, Anastasia Janas, Ujjwal Baid, Divya Ramakrishnan, Leon Jekel, Kiril Krantchev, Harrison Moy, Rachit Saluja, Klara Osenberg, Klara Wilms, Manpreet Kaur, Arman Avesta, Gabriel Cassinelli Pedersen, Nazanin Maleki, Mahdi Salimi, Sarah Merkaj, Marc von Reppert, Niklas Tillmans, Jan Lost, Khaled Bousabarah, Wolfgang Holler, MingDe Lin, Malte Westerhoff, Ryan Maresca, Katherine E. Link, Nourel hoda Tahon, Daniel Marcus, Aristidis Sotiras, Pamela LaMontagne, Strajit Chakrabarty, Oleg Teytelboym, Ayda Youssef, Ayman Nada, Yuri S. Velichko, Nicolo Gennaro, Connectome Students, Group of Annotators, Justin Cramer, Derek R. Johnson, Benjamin Y. M. Kwan, Boyan Petrovic, Satya N. Patro, Lei Wu, Tiffany So, Gerry Thompson, Anthony Kam, Gloria Guzman Perez-Carrillo, Neil Lall, Group of Approvers, Jake Albrecht, Uduanna Anazodo, Marius George Lingaru, Bjoern H Menze, Benedikt Wiestler, Maruf Adewole, Syed Muhammad Anwar, Dominic LaBella, Hongwei Bran Li, Juan Eugenio Iglesias, Keyvan Farahani, James Eddy, Timothy Bergquist, Verena Chung, Russel Takeshi Shinohara, Farouk Dako, Walter Wiggins, Zachary Reitman, Chunhao Wang, Xinyang Liu, Zhifan Jiang, Koen Van Leemput, Marie Piraud, Ivan Ezhov, Elaine Johanson, Zeke Meier, Ariana Familiar, Anahita Fathi Kazerooni, Florian Kofler, Evan Calabrese, Sanjay Aneja, Veronica Chiang, Ichiro Ikuta, Umber Shafique, Fatima Memon, Gian Marco Conte, Spyridon Bakas, Jeffrey Rudie, and Mariam Aboian. The brain tumor segmentation (brats-mets) challenge 2023: Brain metastasis segmentation on pre-treatment mri, 2023.
- Ahmed W Moawad, Anastasia Janas, Ujjwal Baid, Divya Ramakrishnan, Rachit Saluja, Nader Ashraf, Leon Jekel, Raisa Amiruddin, Maruf Adewole, Jake Albrecht, Uduanna Anazodo, Sanjay Aneja, Syed Muhammad Anwar, Timothy Bergquist, Evan Calabrese, Veronica Chiang, Verena Chung, Gian Marco Conte, Farouk Dako, James Eddy, Ivan Ezhov, Ariana Familiar, Keyvan Farahani, Juan Eugenio Iglesias, Zhifan Jiang, Elaine Johanson, Anahita Fathi Kazerooni, Florian Kofler, Kiril Krantchev, Dominic LaBella, Koen Van Leemput, Hongwei Bran Li, Marius George Lingurar, Katherine E Link, Xinyang Liu, Nazanin Maleki, Zeke Meier, Bjoern H Menze, Harrison Moy, Klara Osenberg, Marie Piraud, Zachary Reitman, Russel Takeshi Shinohara, Nourel Hoda Tahon, Ayman Nada, Yuri S Velichko, Chunhao Wang, Benedikt Wiestler, Walter Wiggins, Umber Shafique, Klara Willms, Arman Avesta, Khaled Bousabarah, Satrajit Chakrabarty, Nicolo Gennaro, Wolfgang Holler, Manpreet Kaur, Pamela LaMontagne, Mingde Lin, Jan Lost, Daniel S Marcus, Ryan Maresca, Sarah Merkaj, Ayman Nada, Gabriel Cassinelli Pedersen, Marc von Reppert, Aristidis Sotiras, Oleg Teytelboym, Niklas Tillmans, Malte Westerhoff, Ayda Youssef, Devon Godfrey, Scott Floyd, Andreas Rauschecker, Javier Villanueva-Meyer, Irada Pflüger, Jaeyoung Cho, Martin Bendszus, Gianluca Brugnara, Justin Cramer, Gloria J Guzman Perez-Carrillo, Derek R Johnson, Anthony Kam, Benjamin Yin Ming Kwan, Lillian Lai, Neil U Lall, Fatima Memon, Satya Narayana Patro, Bojan Petrovic, Tiffany Y So, Gerard Thompson, Lei Wu, E Brooke Schrickel, Anu Bansal, Frederik Barkhof, Cristina Besada, Sammy Chu, Jason Druzgal, Alexandru Duso, Luciano Farage, Fabricio Feltrin, Amy Fong, Steve H Fung, R Ian Gray,

- Ichiro Ikuta, Michael Iv, Alida A Postma, Amit Mahajan, David Joyner, Chase Krumpelman, Laurent Letourneau-Guillon, Christie M Lincoln, Mate E Maros, Elka Miller, Fanny Morón, Esther A Nimchinsky, Ozkan Ozsarlak, Uresh Patel, Saurabh Rohatgi, Atin Saha, Anousheh Sayah, Eric D Schwartz, Robert Shih, Mark S Shiroishi, Juan E Small, Manoj Tanwar, Jewels Valerie, Brent D Weinberg, Matthew L White, Robert Young, Vahe M Zohrabian, Aynur Azizova, Melanie Maria Theresa Brüßeler, Pascal Fehringer, Mohanad Ghonim, Mohamed Ghonim, Athanasios Gkampenis, Abdullah Okar, Luca Pasquini, Yasaman Sharifi, Gagandeep Singh, Nico Sollmann, Theodora Soumala, Mahsa Taherzadeh, Nikolay Yordanov, Philipp Vollmuth, Martha Foltyn-Dumitru, Ajay Malhotra, Aly H Abayazeed, Francesco Dellepiane, Philipp Lohmann, Víctor M Pérez-García, Hesham Elhalawani, Sanaria Al-Rubaiey, Rui Duarte Armindo, Kholod Ashraf, Moamen M Asla, Mohamed Badawy, Jeroen Bisschop, Nima Broomand Lomer, Jan Bukatz, Jim Chen, Petra Cimflova, Felix Corr, Alexis Crawley, Lisa Deptula, Tasneem Elakhdar, Islam H Shawali, Shahriar Faghani, Alexandra Frick, Vaibhav Gulati, Muhammad Ammar Haider, Fátima Hierro, Rasmus Holmboe Dahl, Sarah Maria Jacobs, Kuang-Chun Jim Hsieh, Sedat G Kandemirli, Katharina Kersting, Laura Kida, Sofia Kollia, Ioannis Koukoulithras, Xiao Li, Ahmed Abouelatta, Aya Mansour, Ruxandra-Catrinel Maria-Zamfirescu, Marcela Marsiglia, Yohana Sarahi Mateo-Camacho, Mark McArthur, Olivia McDonnell, Maire McHugh, Mana Moassefi, Samah Mostafa Morsi, Alexander Muntenu, Khanak K Nandolia, Syed Raza Naqvi, Yalda Nikanpour, Mostafa Alnoury, Abdullah Mohamed Aly Nouh, Francesca Pappafava, Markand D Patel, Samantha Petrucci, Eric Rawie, Scott Raymond, Borna Roohani, Sadeq Sabouhi, Laura M Sanchez-Garcia, Zoe Shaked, Pokhraj P Suthar, Talissa Altes, Edwin Isufi, Yaseen Dhermesh, Jaime Gass, Jonathan Thacker, Abdul Rahman Tarabishy, Benjamin Turner, Sebastiano Vacca, George K Vilanilam, Daniel Warren, David Weiss, Klara Willms, Fikadu Worede, Sara Yousry, Wondwossen Lerebo, Alejandro Aristizabal, Alexandros Karargyris, Hasan Kassem, Sarthak Pati, Micah Sheller, Spyridon Bakas, Jeffrey D Rudie, and Mariam Aboian. The brain tumor segmentation - metastases (BraTS-METS) challenge 2023: Brain metastasis segmentation on pre-treatment MRI. *ArXiv*, June 2024.
- L.G. Nyul, J.K. Udupa, and Xuan Zhang. New variants of a method of mri scale standardization. *IEEE Transactions on Medical Imaging*, 19(2):143–150, 2000. .
- Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. November 2017. .
- Ashay Patel, Petru-Daniel Tudosiu, Walter HL Pinaya, Mark S Graham, Olusola Adeleke, Gary Cook, Vicky Goh, Sebastien Ourselin, and M Jorge Cardoso. Self-supervised anomaly detection from anomalous training data via iterative latent token masking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2402–2410, 2023.
- Ken Perlin. An image synthesizer. *ACM Siggraph Computer Graphics*, 19(3):287–296, 1985.
- Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. June 2021. .
- Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-anogan: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44, 2019. ISSN 1361-8415. . URL <https://www.sciencedirect.com/science/article/pii/S1361841518302640>.
- Lars Schönlau. *Optimale räumliche Orientierung und Angulierung von T2*- und SWI-Sequenzen im Rahmen eines epilepsiespezifischen MRT-Protokolls*. doctoralthesis, Ruhr-Universität Bochum, Universitätsbibliothek, 2018.
- David Schuhmacher, Stephanie Schörner, Claus Küpper, Frederik Großserueschkamp, Carlo Sternemann, Celine Lugnier, Anna-Lena Kraeft, Hendrik Jütte, Andrea Tannapfel, Anke Reinacher-Schick, Klaus Gerwert, and Axel Mosig. A framework for falsifiable explanations of machine learning models with an application in computational pathology. *Medical image analysis*, 82:102594, August 2022. ISSN 1361-8423. .
- Johannes Schwarz, Lena Will, Jörg Wellmer, and Axel Mosig. A patch-based student-teacher pyramid matching approach to anomaly detection in 3d magnetic resonance imaging. In *Medical Imaging with Deep Learning*, 2024. URL <https://openreview.net/pdf?id=vh01Nd5PC1>.
- Shai Shalev-Shwartz and Shai Ben-David. Understanding machine learning: From theory to algorithms, 2014.
- Md Amran Siddiqui, Alan Fern, Thomas G Dietterich, and Shubhomoy Das. Finite sample complexity of rare pattern anomaly detection. In *UAI*, volume 16, pages 686–695, 2016.
- Jaime Simarro, Ezequiel de la Rosa, Thijs Vande Vyvere, David Robben, and Diana M. Sima. Unsupervised 3d brain anomaly detection. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. BrainLes 2020. Lecture Notes in Computer Science, vol 12658*. Springer, Cham (2021), October 2020. .

- J.G. Sled, A.P. Zijdenbos, and A.C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE Transactions on Medical Imaging*, 17(1):87–97, 1998. .
- Hannah Spitzer, Mathilde Ripart, Kirstie Whitaker, Felice D’Arco, Kshitij Mankad, Andrew A Chen, Antonio Napolitano, Luca De Palma, Alessandro De Benedictis, Stephen Foldes, Zachary Humphreys, Kai Zhang, Wenhan Hu, Jiajie Mo, Marcus Likeman, Shirin Davies, Christopher Güttler, Matteo Lenge, Nathan T Cohen, Yingying Tang, Shan Wang, Aswin Chari, Martin Tisdall, Nuria Bargallo, Estefanía Conde-Blanco, Jose Carlos Pariente, Saül Pascual-Diaz, Ignacio Delgado-Martínez, Carmen Pérez-Enríquez, Ilaria Lagorio, Eugenio Abela, Nandini Mullatti, Jonathan O’Muircheartaigh, Katy Vecchiato, Yawu Liu, Maria Eugenia Caligiuri, Ben Sinclair, Lucy Vivash, Anna Willard, Jothy Kandasamy, Ailsa McLellan, Drahoslav Sokol, Mira Semmelroch, Ane G Kloster, Giske Opheim, Letícia Ribeiro, Clarissa Yasuda, Camilla Rossi-Espagnet, Khalid Hamandi, Anna Tietze, Carmen Barba, Renzo Guerrini, William Davis Gaillard, Xiaozhen You, Irene Wang, Sofía González-Ortiz, Mariasavina Severino, Pasquale Striano, Domenico Tortora, Reetta Kälviäinen, Antonio Gambardella, Angelo Labate, Patricia Desmond, Elaine Lui, Terence O’Brien, Jay Shetty, Graeme Jackson, John S Duncan, Gavin P Winston, Lars H Pinborg, Fernando Cendes, Fabian J Theis, Russell T Shinohara, J Helen Cross, Torsten Baldeweg, Sophie Adler, and Konrad Wagstyl. Interpretable surface-based detection of focal cortical dysplasias: a multi-centre epilepsy lesion detection study. *Brain*, 145(11):3859–3871, November 2022.
- Hannah Spitzer, Mathilde Ripart, Abdulah Fawaz, Logan Z. J. Williams, Emma C. Robinson, Juan Eugenio Iglesias, Sophie Adler, and Konrad Wagstyl. Robust and generalisable segmentation of subtle epilepsy-causing lesions: A graph convolutional approach. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 420–428, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43993-3.
- Tom F Sterkenburg and Peter D Grünwald. The no-free-lunch theorems of supervised learning. *Synthese*, 199(3): 9979–10015, 2021.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Nicholas J. Tustison, Brian B. Avants, Philip A. Cook, Yuanjie Zheng, Alexander Egan, Paul A. Yushkevich, and James C. Gee. N4itk: Improved n3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, 2010. .
- Pierre Vigier. perlin-numpy, 2020. URL <https://github.com/pvigier/perlin-numpy>.
- Lennart Walger, Sophie Adler, Konrad Wagstyl, Leonie Henschel, Bastian David, Valeri Borger, Elke Hattingen, Hartmut Vatter, Christian E. Elger, Torsten Baldeweg, Felix Rosenow, Horst Urbach, Albert Becker, Alexander Radbruch, Rainer Surges, Martin Reuter, Fernando Cendes, Zhong Irene Wang, Hans-Jürgen Huppertz, and Theodor Rüber. Artificial intelligence for the detection of focal cortical dysplasia: Challenges in translating algorithms into clinical practice. *Epilepsia*, 64(5):1093–1112, 2023. . URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/epi.17522>.
- Alan Q Wang, Batuhan K Karaman, Heejong Kim, Jacob Rosenthal, Rachit Saluja, Sean I Young, and Mert R Sabuncu. A framework for interpretability in machine learning for medical imaging. *IEEE Access*, 12:53277–53292, 2024a.
- Alan Q Wang, Rachit Saluja, Heejong Kim, Xinzi He, Adrian Dalca, and Mert R Sabuncu. Brainmorph: A foundational keypoint model for robust and flexible brain mri registration. *arXiv preprint arXiv:2405.14019*, 2024b.
- Kevin Y. Wang, Oluwatoyin R. Idowu, and Doris D.M. Lin. Chapter 24 - radiology and imaging for cavernous malformations. In Robert F. Spetzler, Karam Moon, and Rami O. Almefty, editors, *Arteriovenous and Cavernous Malformations*, volume 143 of *Handbook of Clinical Neurology*, pages 249–266. Elsevier, 2017. . URL <https://www.sciencedirect.com/science/article/pii/B9780444636409000242>.
- Mingqing Wang, Jiawei Li, Zhenyang Li, Chengxiao Luo, Bin Chen, Shu-Tao Xia, and Zhi Wang. Unsupervised anomaly detection with local-sensitive vqvae and global-sensitive transformers. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1080–1084. IEEE, 2023.
- Jörg Wellmer, Carlos M. Quesada, Lars Rothe, Christian E. Elger, Christian G. Bien, and Horst Urbach. Proposal for a magnetic resonance imaging protocol for the detection of epileptogenic lesions at early outpatient stages. *Epilepsia*, 54(11):1977–1987, 2013. . URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/epi.12375>.

David H Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.

Julian Wyatt, Adam Leach, Sebastian M. Schmon, and Chris G. Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 649–655, 2022. .

Ziv Yaniv, Bradley C. Lowekamp, Hans J. Johnson, and Richard Beare. SimplelTK image-analysis notebooks: a collaborative environment for education and reproducible research. *Journal of Digital Imaging*, 31(3):290–303, Jun 2018. ISSN 1618-727X. . URL <https://doi.org/10.1007/s10278-017-0037-8>.

Appendix A. Anomaly Detection and Probably Approximately Correct Learning

The theory of machine learning has brought forth the concept of *probably approximately correct* (PAC) learning that helps to understand the characteristics and limitations of different learning algorithms. Yet, PAC learning addresses supervised learning, raising the question in what manner results regarding PAC learning relate to anomaly detection approaches. We here provide a brief and simple, but formal consideration under what circumstances an anomaly detection algorithm can be considered an instance of PAC learning.

We first provide a brief summary of PAC learnability following Shalev-Shwartz and Ben-David (2014). PAC learning deals with identifying models h (also referred to as hypothesis) in the form of a mathematical function $h: \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X}$ is the set of all data points, also referred to as the data domain, and $\mathcal{Y}$ is the label domain. In the context of anomaly detection, considering the case of binary labels $\mathcal{Y} = \{0, 1\}$ will be fully sufficient. A *learning algorithm* $\mathcal{A}$ maps a finite set of training data $T \subset \mathcal{X} \times \mathcal{Y}$ to a model $h = \mathcal{A}(T)$. We assume that the training data are drawn from a probability distribution $\mathcal{D}$ over $\mathcal{X}$, and that for $(x, y) \in T$, $y = g(x)$, where g is the ground truth function unknown to the learning algorithm.

The outcome $h = \mathcal{A}(T)$ of a learning algorithm is assessed by measuring the generalization error as the probability of h yielding the ground truth result for an x randomly sampled from $\mathcal{X}$ according to distribution $\mathcal{D}$, formally

$$\mathcal{L}_{\mathcal{D},g}(h) = \mathbf{P}_{x \sim \mathcal{D}}(h(x) \neq g(x)).$$

PAC learning deals with questions regarding the number of training data points that are required to learn certain classes of functions. Such a class of functions, also referred to as a hypothesis space, is a set

$$\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}},$$

where $\mathcal{Y}^{\mathcal{X}}$ denotes the set of all functions that map from $\mathcal{X}$ to $\mathcal{Y}$.

Given a hypothesis class $\mathcal{H}$, we can now ask whether a learning algorithm whose outcome is restricted to models from $\mathcal{H}$ can guarantee accomplishing a generalization error of at most ε for arbitrary ground truth functions g , given a sufficient number of training examples. Since the generalization error varies under different training data, we will deem it sufficient if the generalization error ε is accomplished with a certain probability δ .

We can now define PAC-learnability of a hypothesis class $\mathcal{H}$ formally as follows:

Definition 1 (PAC learnable)

A hypothesis space $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is PAC-learnable iff there is

a learning algorithm $\mathcal{A}$ such that, for any distribution $\mathcal{D}$ over $\mathcal{X}$ and for an arbitrary generalization error threshold ε and for any probability δ , there is an integer $m := m_{\mathcal{H}}(\varepsilon, \delta)$ with the property that for any training data set T satisfying $|T| \geq m$, with probability at least $1 - \delta$, $h = \mathcal{A}(T) \in \mathcal{H}$ will satisfy $\mathcal{L}_{h,\mathcal{D}} \leq \varepsilon$.

A key theorem in computational learning theory is the no-free-lunch theorem:

Theorem 2 (No-Free-Lunch Theorem for Supervised Learning)

The hypothesis space $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$ of all functions mapping from $\mathcal{X}$ to $\mathcal{Y}$ is not PAC-learnable.

The no-free-lunch theorem has important consequences: The hypothesis space $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$ would be the only basis for a learning algorithm to be unbiased, as no mathematical function would be a priori excluded by the algorithm before being confronted with a training dataset. The no-free-lunch theorem states that this is impossible, so that every learning algorithm must be biased by what is usually referred to as its inductive bias, i.e., the subset of $\mathcal{Y}^{\mathcal{X}}$ which it is capable of learning.

It is important to note that the notion of PAC-learnable does not cover anomaly detection, specifically semi-supervised anomaly detection: here, the training data are sampled from

$$\mathcal{X}_0 := \{x \in \mathcal{X} \mid g(x) = 0\} \quad (1)$$

rather than from $\mathcal{X}$.

Since the notion of PAC-learnable does not cover semi-supervised anomaly detection, the no-free-lunch theorem does not cover anomaly detection either. In order to make claims about the learnability of a hypothesis, we can adjust the notion of PAC-learnability to anomaly detection as follows:

Definition 3 (PAC Anomaly Learnable)

A hypothesis

space $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is PAC-anomaly-learnable iff there is a learning algorithm $\mathcal{A}$ such that, for any distribution $\mathcal{D}_0$ over $\mathcal{X}_0$ and for an arbitrary generalization error threshold ε and for any probability δ , there is an integer $m := m_{\mathcal{H}}(\varepsilon, \delta)$ with the property that for any training data set T satisfying $|T| \geq m$, with probability at least $1 - \delta$, $h = \mathcal{A}(T) \in \mathcal{H}$ will satisfy $\mathcal{L}_{h,\mathcal{D}} \leq \varepsilon$.

The difference between PAC-learnability and PAC-anomaly learnability is small, but significant. In PAC-anomaly learning, the training data are sampled from the subset of all negatively-labelled points in the domain $\mathcal{D}$ rather than from the complete domain, as in the case of PAC-learning. As a consequence, it becomes difficult to reduce PAC-learnability to PAC anomaly learnability, which would imply the no-free-lunch theorem for PAC-anomaly learnability. Intuitively,

however, it seems that PAC-anomaly learnability must be much harder than PAC learnability, as training relies on negative data points only. We hence formulate the following, yet unproven conjecture:

Conjecture 4 (No-Free-Lunch Theorem for Anomaly Detection)

The hypothesis space $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$ of all functions mapping from $\mathcal{X}$ to $\mathcal{Y}$ is not PAC-anomaly learnable.

While Conjecture 4 remains without proof for the time being, our manuscript is based on the assumption that no-free-lunch theorems do apply to anomaly detection: if there are "no free lunches" in anomaly detection, the consequences for anomaly detection are the same as for supervised learning, namely that each anomaly detection method must inevitably have an inductive bias.

Appendix B. Volume-stratified evaluation of tumor metastasis lesion in the BraTS-METS dataset

Experiment

In order to stratify the performance of lesion detection on the BraTS-METS dataset by lesion volume, we performed validation at the object detection (*ObD*) level. Specifically, we divided lesions into six bins by volume (0–10, 10–50, 50–100, 100–500, 500–1000, and 1000+ mm³). A lesion is counted as detected if any of the top- k anomaly peaks falls within radius r of its center (for point annotations) or overlaps with its segmentation mask (for volumetric ground truth).

Results

Supplementary Table 3 presents volume-stratified detection results comparing *SpyCAT* against β -VAE.

Discussion

The volume-stratified analysis reveals clear performance trends across lesion sizes for both methods. *SpyCAT* demonstrates nearly twofold higher overall sensitivity (19.4% vs 10.4%), detecting 289 out of 1487 lesions compared to 154 for β -VAE. *SpyCAT* achieves substantially higher detection sensitivity across all volume categories compared to β -VAE. The advantage is more pronounced for small lesions and decreases with lesion size: In the range between 10 and 500 mm³, *SpyCAT* identifies more than five times as many lesions as β -VAE; for lesions larger than 500 mm³, *SpyCAT* identifies 1.57 times more lesions than β -VAE.

The degradation in performance towards smaller lesions is shared by both approaches, and is well in line with a similar degradation observed for supervised methods in

Moawad et al. (2023). This pattern strongly confirms the inherent challenge of detecting very small anomalies in high-resolution 3D volumes where local context may be insufficient for differentiation from normal anatomical variation. While the absolute detection rates for small lesions appear disappointing at first, they must be seen in relation to the fully external validation setting, where both models were trained exclusively on the independent IXI dataset without any exposure to BraTS-METS data or metastatic lesions during training.

Appendix C. Volume-stratified cross-domain evaluation on the MSLesSeg dataset

Experiment

In order to further assess our method's generalizability across different pathologies and scanner protocols, we conducted additional evaluations on the multiple sclerosis (MS) lesions from the MSLesSeg (Guarnera et al., 2025) dataset.

The test dataset of the MSLesSeg dataset (Guarnera et al., 2025) provides 22 MRI studies. Since it is well-established that the FLAIR sequence is particularly effective at highlighting MS lesions, we only utilized the FLAIR sequence and disregarded all other MRI sequences. For training, we used 304 non-lesional FLAIR sequences from our AIM.E dataset (noting that the IXI dataset used for the BraTS datasets does not provide FLAIR sequences). The MSLesSeg dataset was split equally into validation and test sets with 11 studies each.

Both the AIM.E-based training data and the MSLesSeg data were processed *as is*, i.e., without any adjustment in the preprocessing procedures, which differ substantially between the two datasets.

Results

We report the same complementary detection metrics stratified by lesion volume as for the BraTS-METS dataset, including *per-image* (*ImLC*) detection and *per-lesion* detection (*ObD*). Image level classification (*ImLC*) of an MRI was considered to be correctly identified as lesional if at least one lesion was detected correctly. At this image level, *SpyCAT* succeeded in 10/11 (90.9%) of all MRIs.

Supplementary Table 4 presents volume-stratified detection results for object detection in the MSLesSeg dataset.

Discussion

The volume-stratified analysis reveals that for large lesions, the performance of *SpyCAT* on the MSLesSeg dataset is similar to the BraTS-METS dataset. *SpyCAT* achieves strong detection for lesions exceeding 1000 mm³ (approximately 50% sensitivity). For lesions smaller than 100 mm³,

Table 3: Volume-stratified per-lesion detection performance comparing *SpyCAT* and β -VAE on the BraTS-METS dataset. Lesions are stratified by volume and evaluated individually. Detection criterion: top- $k = 10$ anomaly peaks with segmentation overlap. Total lesion count: 1487.

Volume (mm ³)	SpyCAT		β -VAE	
	Detected/Total	Sensitivity	Detected/Total	Sensitivity
0–10	0/60	0.0%	0/60	0.0%
10–50	5/141	3.5%	0/141	0.0%
50–100	10/174	5.7%	2/174	1.1%
100–500	41/517	7.9%	4/517	0.8%
500–1000	5/107	4.7%	2/107	1.9%
1000+	228/488	46.7%	146/488	29.9%
Overall	289/1487	19.4%	154/1487	10.4%

Table 4: Volume-stratified lesion detection on the MSLesSeg (MS lesions) dataset. Each lesion is evaluated individually. Detection criterion: top- $k = 10$ anomaly peaks, $r = 9$ mm radius.

Volume (mm ³)	MSLesSeg (MS)	
	Detected/Total	Sensitivity
0–10	0/10	0.0%
10–50	1/10	10.0%
50–100	0/15	0.0%
100–500	13/102	12.7%
500–1000	12/52	23.1%
1000+	22/44	50.0%
Overall	48/233	20.6%

Appendix D. Visualization of point annotations in the AIM.E cavernoma dataset

Figure 5 displays the complete set of 22 cavernomas used in this study. Each image represents a coronal MRI slice with volume-based min-max intensity, centered at the expert-annotated lesion location. The red circles indicate the lesion positions and serve solely for visualization purposes—they do not reflect actual lesion sizes, which vary considerably across cases. All images are point-annotated 2D coronal views extracted from SWI-weighted sequences.

however, performance degrades substantially. This degradation is stronger than for the BraTS-METS dataset. This is likely due, at least in part, to the lack of adjustment in pre-processing between the AIM.E-based training dataset and the MSLesSeg validation and test dataset. The IXI-based training data and the BraTS-METS validation and test data underlying Supplement II were much better aligned in terms of preprocessing than for the MS lesion detection shown here.

The notably higher per-image detection rates (90.9% for MSLesSeg, 72.4% for BraTS-METS) compared to per-lesion rates demonstrate that *SpyCAT* successfully flags studies containing lesions, even when individual small lesions go undetected. While the degradation of model performance towards smaller lesions seems disappointing at first, it needs to be seen in reference to other anomaly detection methods, rather than in contrast to supervised methods.

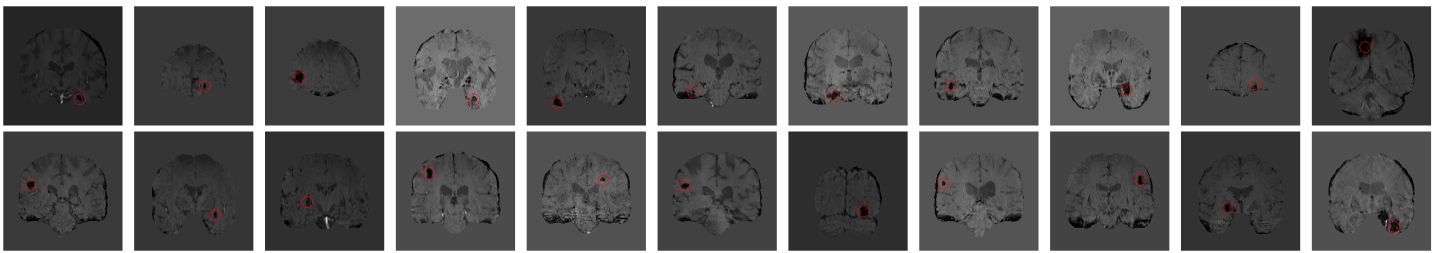


Figure 5: Complete dataset of 22 cavernomas with point annotations. Each panel shows a coronal MRI slice with volume-based min-max normalization centered at the lesion location marked with a red circle. The circles indicate lesion positions for visualization only and do not represent actual lesion dimensions. Images are arranged in a grid format showing the anatomical diversity of cavernoma presentations across different patients and scanner protocols.