Deep Generative Models for Unsupervised Anomaly Detection in Magnetic Resonance Imaging of the Brain

Vom Promotionsausschuss der
Technischen Universität Hamburg

zur Erlangung des akademischen Grades

Doktor-Ingenieur (Dr.-Ing.)

genehmigte kumulative Dissertation

von
Finn Tobias Behrendt

aus
Hamburg

2026

Gutachter:
Prof. Dr.-Ing. Alexander Schlaefer
Prof. Dr. Mattias Heinrich

Datum der mündlichen Prüfung:
04.12.2025

Contents

List of Figures

List of Tables

Abstract

Deep learning is progressively explored for medical image analysis, offering the potential to assist in detecting abnormalities. However, most approaches rely on supervised learning, which requires large annotated data sets and is inherently limited to known pathologies. Unsupervised anomaly detection provides an alternative by learning a distribution of healthy anatomy, allowing deviations from this distribution to be identified as potential anomalies without requiring pathology-specific annotations. A common strategy in brain MRI analysis is to use generative models to reconstruct healthy brain anatomy from input images. Anomalies are then identified based on reconstruction errors, i.e., regions where the model fails to reconstruct the input. A key challenge in this process is ensuring that the model reconstructs only healthy anatomy rather than replicating pathological structures. This balance depends on the choice and strength of regularization techniques applied during the reconstruction process.

In this thesis, we incorporate additional contextual information into both the reconstruction and anomaly scoring processes. We adapt variational autoencoders and diffusion models to improve reconstruction quality while mitigating the risk of replicating abnormal structures. Furthermore, we refine anomaly scoring mechanisms to more reliably differentiate pathological deviations from reconstruction artifacts.

We evaluate our approaches on diverse data sets covering various brain pathologies. Our results demonstrate improvements in reconstruction quality and anomaly detection performance, outperforming state-of-the-art methods. Furthermore, we show that our approaches can be combined with supervised models, enabling effective detection of known pathologies while maintaining the ability to generalize to previously unseen. These advancements contribute towards clinically relevant anomaly detection for brain MRI analysis.

1 Introduction

Magnetic resonance imaging (MRI) is a non-invasive diagnostic tool widely used in neurology. Its ability to capture detailed anatomical and pathological information makes it valuable for detecting neurological disorders [1]. However, interpreting brain MRI scans is a complex task that requires substantial expertise. Radiologists must assess numerous anatomical structures and potential abnormalities in each scan. This process is time-intensive and error-prone [2, 3, 4] as also indicated by high inter-rater variabilities observed in clinical studies [5, 6].

Deep learning has emerged as a powerful tool for medical image analysis, offering a way to assist radiologists in interpreting imaging data [7, 8]. It has been successfully applied across a wide range of applications [9, 10, 11], with a level of precision comparable to that of human experts [12, 13, 14]. Therefore, deep learning models show promise in supporting radiologists in interpreting brain MRI scans. However, most methods rely predominantly on supervised learning, which requires large annotated data sets that are costly and time-consuming to obtain [15, 16]. More critically, supervised models are constrained by their training data, limiting their ability to detect novel or rare pathologies absent from the training distribution [17]. This limitation is especially problematic for rare diseases, where data is often insufficient to build extensive data sets [18]. Moreover, for clinical applications, it is imperative to detect any anomaly, including secondary or incidental findings that may not be the primary focus of the examination but could indicate serious pathologies [19].
Unsupervised anomaly detection (UAD) offers a promising alternative by reducing the reliance on annotated data. Instead of directly mapping images to anomaly labels, UAD models learn a reference distribution of healthy brain anatomy. Deviations from this distribution are flagged as anomalies, enabling the detection of previously unseen pathologies. A prevalent approach within UAD employs generative models (GM) trained to reconstruct healthy brain MRI scans. At test time, it is assumed that the GM fails to reconstruct abnormal structures unseen during training. Anomalies are then identified by comparing the original image to its reconstruction (see Figure 1.1). This approach has the potential to detect and localize a wide variety of anomalies without requiring pathology-specific annotations.
Despite its promise, this reconstruction-based approach faces critical challenges. The reconstruction process must achieve high accuracy for healthy structures while avoiding the trivial replication of input images. To this end, regularization mechanisms such as latent-space bottlenecks in autoencoders or denoising objectives in diffusion models are employed. However, these regularization strategies introduce a trade-off, as illustrated in Figure 1.2. Insufficient regularization can result in replicating both normal and abnormal structures, hindering anomaly detection (see 'under-regularization' in Figure 1.2). Conversely, excessive regularization may degrade reconstruction quality, introducing artifacts that resemble anomalies (see 'over-regularization' in Figure 1.2).
Moreover, even with optimal regularization, GMs may struggle to perfectly reconstruct

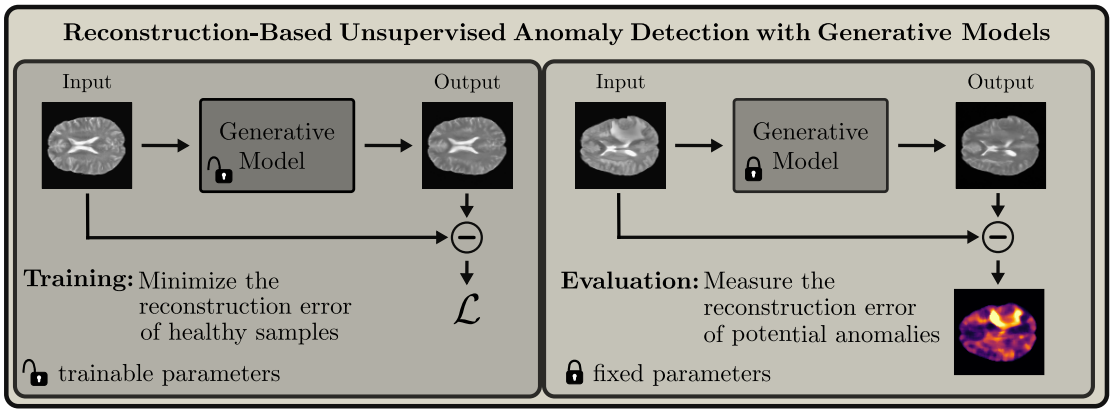


Fig. 1.1: Unsupervised anomaly detection using generative models. Left: The generative model learns to reconstruct healthy brain MRI scans during training. Right: The anomalies are identified as deviations from the learned distribution of healthy anatomy at test time. The anomaly map highlights regions with the highest reconstruction errors, indicating potential abnormalities.

healthy structures ('balanced regularization' in Figure 1.2). As a result, robust post-processing and anomaly scoring mechanisms are essential to distinguish genuine anomalies from reconstruction artifacts.

In this thesis, we investigate whether incorporating supplementary contextual information into the reconstruction process and anomaly scoring mechanism can address the challenge of balancing regularization and reconstruction accuracy. We hypothesize that contextual data, such as three-dimensional spatial information, surrounding regions of erased patches, or abstract features derived from the input image, could provide additional guidance during reconstruction. This information could enable more coherent reconstructions of healthy structures while maintaining sufficient regularization to avoid trivial solutions, such as merely replicating the input image.

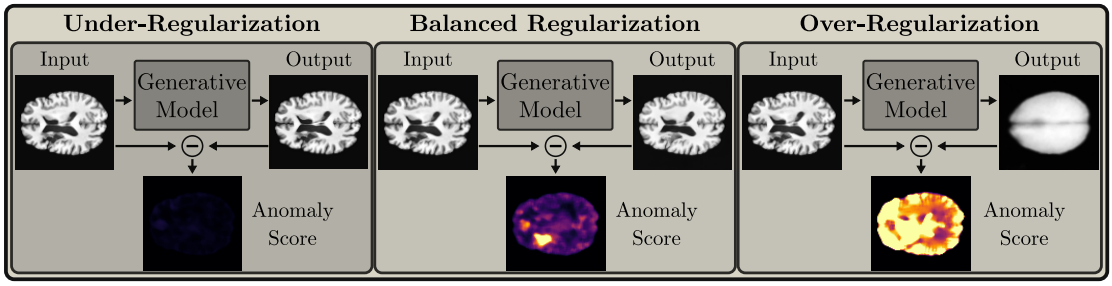


Fig. 1.2: Trade-off between reconstruction quality and regularization. Left: 'Under-Regularization' can lead to the replication of both normal and abnormal structures. Middle: 'Balanced Regularization' enables high reconstruction accuracy for healthy structures without replicating unhealthy ones. Right: 'Over-Regularization' can degrade reconstruction quality, introducing artifacts that resemble anomalies. The generative model used for visualizations is a diffusion model, where varying noise levels are introduced during reconstruction to simulate different degrees of regularization.

For anomaly scoring, additional context could capture structural relationships from neighboring regions or the variability of pixels across multiple reconstructions. We hypothesize that integrating this context into anomaly scoring could help to distinguish genuine anomalies from reconstruction artifacts, enhancing the segmentation performance. These hypotheses lead to two principal research questions:

1. **Reconstruction Quality** *Can incorporating supplementary contextual information improve the accuracy of reconstructions without compromising regularization?*

2. **Anomaly Scoring** *Can incorporating supplementary contextual information enhance the segmentation of anomalies in reconstruction-based anomaly scoring?*

In this thesis, we address these questions by introducing advanced GMs and novel anomaly scoring mechanisms:
Advancing Generative Models: We introduce regularized three-dimensional Variational Autoencoders [20] and incorporate additional context into denoising diffusion probabilistic models (DDPMs) [21, 22] to improve the coherence and accuracy of reconstructions. These advancements address the challenge of balancing reconstruction quality with regularization, providing robust backbones for reconstruction-based UAD in brain MRI.
Novel Anomaly Scoring Mechanisms: We propose new methods for anomaly scoring, including an ensemble approach using structural similarity across multiple scales [23] and a Mahalanobis Distance-based approach that leverages the probabilistic nature of DDPMs, accounting for the variability of multiple reconstructions [24]. Moreover, we propose a framework that combines unsupervised GMs with supervised anomaly scoring to improve the detection of known anomalies while enhancing generalization to unknown ones [25].
Our findings indicate that the proposed mechanisms to integrate information from the input image enable enhanced reconstruction quality while ensuring sufficient regularization to prevent the replication of unhealthy structures. Furthermore, advanced anomaly scoring techniques, which account for structural differences and the normal variability of reconstructions, improve the segmentation performance. The proposed framework combining GMs with supervised scoring highlights the potential of hybrid UAD methods for clinical applications. Overall, our approaches significantly enhance reconstruction quality and anomaly detection performance, advancing the state-of-the-art in UAD for brain MRI.

This thesis is structured as follows: Chapter 2 provides an overview of the relevant background for this thesis. Chapter 3 describes the methodology of the proposed models and scoring mechanisms, including the data sets, metrics, and experimental setups. Chapter 4 presents the results of our experiments, including an assessment of the proposed methods on various data sets and a comparison with the state-of-the-art. Chapter 5 discusses the findings in the context of existing literature, highlights the limitations of our methods and UAD in general, and proposes directions for future research. Chapter 6 concludes the thesis by summarizing key contributions, limitations, and outlook. Chapter 7 offers a summary, and Chapter 8 provides the publications related to this thesis.

2 Background

In this chapter, we introduce the basic concepts of MRI and its acquisition process. Furthermore, we present the fundamental principles of machine learning, deep learning, and GMs. Lastly, we introduce the concept of anomaly detection.

2.1 Magnetic Resonance Imaging

This section is based on the book of Weishaupt et al. [26].
The fundamental principle of MRI is based on measuring the magnetic properties of hydrogen nuclei in the tissue of interest and their interaction with external magnetic fields. The protons of the electrically neutral hydrogen atoms possess an angular momentum, referred to as spin. Due to its rotating electric charge, the proton generates a magnetic moment B, which can be aligned in the direction of an external magnetic field B_0. Meanwhile, nuclei undergo a precession process with a frequency proportional to the external magnetic field B_0. The precession frequency is called the Larmor frequency and is calculated using the Larmor equation

$$\omega_0 = \gamma_0 \cdot B_0. \tag{2.1}$$

Here, γ_0 is the gyromagnetic ratio of the hydrogen nuclei, which is a tissue-dependent constant. When a strong external magnetic field, B_0, is applied for a sufficient duration, the spins attain a state of stability in the direction of the magnetic field, resulting in parallel and antiparallel alignment. The difference in the number of spins in the parallel and antiparallel alignment enables the generation of a longitudinal net magnetization vector M_Z in the direction of the magnetic field (See Figure 2.1). If an electromagnetic wave with a frequency corresponding to the Larmor frequency is induced by a Radio Frequency (RF) pulse, the spins are deflected, which causes the net magnetization M_Z to decrease. Concurrently, a transverse magnetization M_{XY} builds up. The angle of the spins relative to the magnetic field is referred to as the flip angle, which can be controlled by modifying the duration and power of the RF pulse. The resulting transversal magnetization M_{XY} generates a signal that can be measured as oscillating voltage by the receiver coils of the MRI scanner. Following the deactivation of the RF pulse, two relaxation processes occur: the spin-lattice relaxation or T1-relaxation and the spin-spin relaxation or T2-relaxation. Both relaxation mechanisms result in the decay of the transverse magnetization M_{XY} and the recovery of the longitudinal magnetization M_Z. The T1-relaxation describes the orientation of the spins along the external magnetic field B_0 due to the emission of energy to the surrounding environment. The resulting increase of the longitudinal magnetization M_Z is described by

$$M_Z(t) = M_0 \cdot (1 - e^{-\frac{t}{T1}}), \tag{2.2}$$

where $T1$ is a tissue-dependent time constant and M_0 is the equilibrium magnetization. The T2-relaxation describes the dephasing of spins resulting from their interaction with

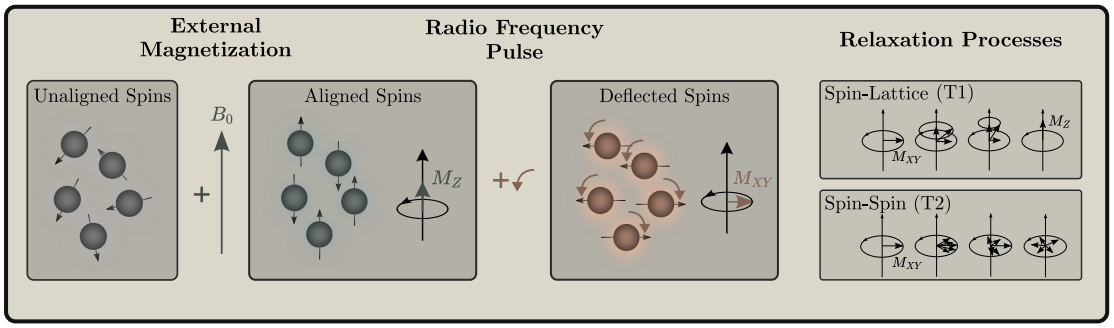


Fig. 2.1: MRI physics and acquisition process, adapted from [26]. From left to right, the alignment of the spins in the external magnetic field B_0, the excitation of the spins by a radio frequency pulse and the relaxation processes are illustrated.

one another and field inhomogeneities. The reduction in the transverse magnetization M_{XY} is described by

$$M_{XY}(t) = M_0 \cdot e^{-\frac{t}{T2}}, \tag{2.3}$$

where $T2$ is also a tissue-dependent time constant. Notably, T1- and T2-relaxation are independent of each other and can co-occur. In most cases, the decay of the transversal magnetization M_{XY} caused by the T2-relaxation is faster than the recovery of the longitudinal magnetization M_Z caused by the T1-relaxation.

MRI Acquisition

To obtain an interpretable MRI scan, it is necessary to encode the signal of the transversal magnetization, as measured by the receiver coils, in a spatial format. This encoding is accomplished by applying magnetic field gradients in the x, y, and z-directions. In the z-direction, a gradient is applied to the external magnetic field B_0 for slice selection. The gradient causes the Larmor frequency to change locally along the z-direction, thereby enabling the excitation of protons in selected slices of the scanned subject. Within a given slice, the spatial encoding is achieved by applying additional magnetic field gradients in the x and y directions. For the y direction, a phase-encoding gradient is used for a limited duration, resulting in a phase shift of the spins in the y direction. For the x direction, a static gradient is applied, leading to a frequency shift of the spins in the x direction. The resulting measurements are organized in a matrix of data points, each representing a unique combination of frequency and phase shift. This matrix is called K-space, where frequency encoding is arranged along the horizontal axis and phase encoding along the vertical axis. Once the K-space is sampled, an inverse Fourier transform is applied to convert the encoded frequency and phase information into an interpretable image.

MRI Contrast

While the proton density determines the maximum possible signal strength, tissue contrast in MRI is influenced by the differences in T1 and T2 relaxation times. Consequently, contrast enhancement can be achieved by using different acquisition sequences, including T1-weighted and T2-weighted sequences. In T1-weighted MRI scans, the repetition time, defined as the interval between two RF pulses, and the echo time, defined as

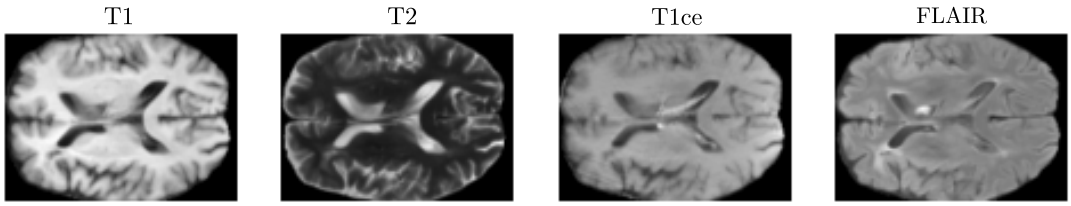


Fig. 2.2: MRI contrast types from the BRATS dataset [28]. From left to right, exemplary T1-weighted, T2-weighted, contrast-enhanced T1-weighted, and FLAIR scans are presented.

the time between the RF pulse and the signal acquisition, are selected to be relatively short. Consequently, tissue types exhibiting a short T1-relaxation time demonstrate increased signal intensity as the longitudinal magnetization M_Z recovers faster. In contrast, tissue types with a long T1-relaxation time exhibit decreased signal intensity as the longitudinal magnetization M_Z recovers slower. Examples of tissue types with short and long T1-relaxation times are fat and water, respectively.
In contrast, T2-weighted MRI scans employ long repetition and echo times. Tissue types exhibiting a long T2-relaxation time demonstrate increased signal intensity as the transverse magnetization M_{XY} decays more slowly. In contrast, tissue types with a short T2-relaxation time exhibit decreased signal intensity as the transverse magnetization M_{XY} decays more rapidly. Exemplary tissue types with long and short T2-relaxation times are water and fat, respectively.
In addition to T1- and T2-weighted MRI scans, other weighting sequences are available, including FLuid-Attenuated Inversion Recovery (FLAIR) MRI scans. FLAIR MRI scans suppress the signal of cerebrospinal fluid and can be beneficial to visualize subtle pathological changes, e.g., induced by multiple sclerosis lesions [27]. Additionally, contrast agents such as gadolinium can enhance the contrast of specific tissues in MRI scans. Figure 2.2 shows an example of a T1-weighted, a contrast-enhanced T1-weighted, a T2-weighted and a FLAIR-weighted MRI scan of an exemplary brain scan.

2.2 Machine Learning

In this section, we provide a brief introduction to machine learning, primarily based on the book of Goodfellow et al. [29].
Machine learning is a subfield of artificial intelligence that focuses on developing algorithms capable of making predictions or decisions based on data. The field is frequently divided into three categories: supervised, unsupervised, and reinforcement learning. To maintain the focus of this thesis, we limit our discussion to supervised and unsupervised learning.

2.2.1 Supervised Learning

In supervised learning, training is performed with a labeled dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, which comprises N input-output pairs, represented as x_i and y_i. The objective is to learn the parameters θ of a function f_θ, which maps the input data (also called features) x_i to the output (also called targets) y_i. The mapping function $\hat{y}_i = f_\theta(x_i)$ can then be

employed to make predictions on new, previously unseen features. The training process involves adjusting the parameters θ to minimize the discrepancy between the predicted output $\hat{y}_i$ and the actual output y_i. Common supervised learning tasks are regression and classification. In regression, the label $y \in \mathbb{R}$ is a continuous value, and the loss function is typically the Mean Squared Error (MSE, $l2$-error) or Mean Absolute Error (MAE, $l1$-error). In the context of classification, the output is a discrete value $\boldsymbol{y}_i \in \mathbb{R}^C$, where C is the number of classes. The loss function is typically the cross-entropy loss, a measure of discrepancy between the predicted probability distribution and the actual distribution of the classes.

2.2.2 Unsupervised Learning

In unsupervised learning, the training process is performed on an unlabeled dataset $\mathcal{D} = \{(x_i)\}_{i=1}^N$ of N input data points x_i. The objective is to model the underlying structure of the training data set $\mathcal{D}$. In many cases, a latent representation of the data is learned that captures essential information. This representation can then be employed for further analysis or to make predictions on new, previously unseen data. A prominent approach is data compression via dimensionality reduction. The objective is to reduce the number of features while preserving the most important information of the data. Common dimensionality reduction algorithms are autoencoders that can be seen as a non-linear generalization of principal component analysis [30, 31]. Other algorithms include t-distributed stochastic neighbor embedding [32] and uniform manifold approximation and projection [33].

2.2.3 Artificial Neural Networks

Artificial Neural Networks (ANNs) are widely adopted machine learning models inspired by the cortical neural networks of the human brain. ANNs typically consist of multiple layers of Artificial Neurons (ANs) connected by weighted edges. In the following, we introduce the basic concepts of ANNs, based on the book of Hagan et al. [34].

Artificial Neuron

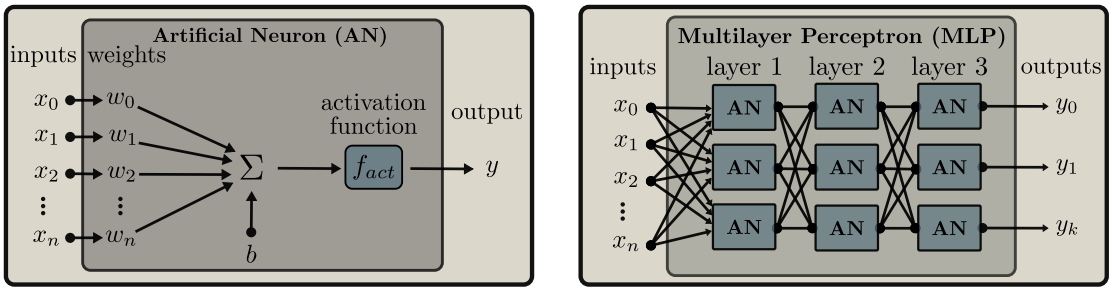
An AN is the basic building block of ANNs. It receives n inputs $\boldsymbol{x} \in \mathbb{R}^n$ and computes a single output $y \in \mathbb{R}$ by weighting the inputs with weights $\boldsymbol{w} \in \mathbb{R}^n$ and calculating the weighted sum of the inputs. After adding a bias term $b \in \mathbb{R}$, the weighted sum is passed through an activation function f_{act}. In vector form, the output y of an AN is given by

$$y = f_{act}(\boldsymbol{w}^T \boldsymbol{x} + b), \tag{2.4}$$

where $\boldsymbol{w}^T$ is the transpose of the weight vector $\boldsymbol{w}$. To learn non-linear mappings, the activation function f_{act} is often chosen to be non-linear. A common activation function is the Rectified Linear Unit (ReLU) function [35], defined as $f_{ReLU}(x) = \max(0, x)$.

Multilayer Perceptron

A Multilayer Perceptron (MLP) consisting of multiple ANs is illustrated in Figure 2.3. An MLP consists of an input layer, one or multiple hidden layers, and an output layer. The input layer receives n inputs $\boldsymbol{x} \in \mathbb{R}^n$ and passes them to the hidden layers. Each hidden layer consists of multiple ANs that compute the weighted sum of the inputs,



(a) Artificial Neuron.

(b) Multilayer Perceptron.

Fig. 2.3: Artificial neuron (left) and multilayer perceptron (right), adapted from [34]. The artificial neuron receives n inputs and computes one output y. The multilayer perceptron consists of multiple layers of artificial neurons. The input layer receives n inputs $\boldsymbol{x}$ and passes them to the hidden layers. The output layer computes k outputs $\boldsymbol{y}$.

add a bias term, and pass the result through an activation function. The output layer computes the final output of the MLP. Compared to the single output in ANs, MLPs can have k outputs $\boldsymbol{y} \in \mathbb{R}^k$. For an MLP with L layers, the output $\boldsymbol{y}$ is computed by

$$\boldsymbol{y} = f_l(f_{l-1}(\dots f_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\dots + \boldsymbol{b}_{l-1}) + \boldsymbol{b}_l), \tag{2.5}$$

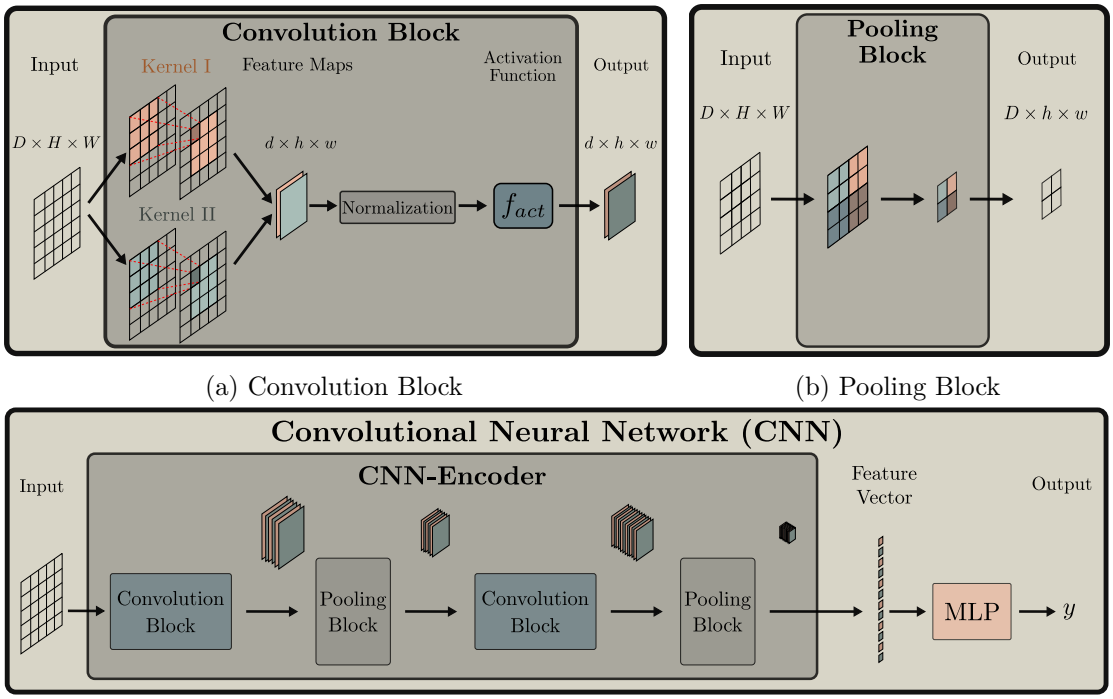
where $\boldsymbol{W}_l$ is the weight matrix of layer l and $\boldsymbol{b}_l$ is the corresponding bias vector. Each weight matrix $\boldsymbol{W}_l \in \mathbb{R}^{d_l \times d_{l-1}}$ connects the d_{l-1} neurons of the previous layer to the d_l neurons of the current layer.

2.3 Deep Learning

Deep learning represents a subfield of machine learning focused on developing deep neural networks comprising multiple hidden layers. In the following, we introduce basic deep learning concepts, concentrating on convolutional neural networks and their applications in image processing.

2.3.1 Convolutional Neural Networks

MLPs that consist of multiple hidden layers are referred to as deep neural networks. While these networks can be applied to various tasks, they are not well-suited for handling grid-like data, such as images. In MLPs, each input feature is connected to every AN in the subsequent layer. These connections result in an exponential increase in the number of parameters as the number of input features increases. Furthermore, MLPs necessitate the flattening of input data, which results in the loss of spatial information. Convolutional Neural Networks (CNN) overcome these constraints, enabling the efficient processing of images while preserving spatial patterns.
The subsequent introduction of CNNs is based on the book of Goodfellow et al. [29]. CNNs consist of multiple layer types, including convolutional, pooling and fully connected layers. Convolutional layers are employed to extract features, whereas pooling layers are used to reduce the spatial dimensions of the data, thereby facilitating the construction of abstract features. The Fully Connected (FC) layers are MLPs that are utilized to make



(a) Convolution Block

(b) Pooling Block

(c) Convolutional Neural Network (CNN)

Fig. 2.4: Illustration of the convolution and pooling blocks (top) of a simple CNN architecture (bottom). The activation functions and normalization layers are included in the convolution block, and the batch dimension is omitted for simplicity.

predictions based on the extracted, lower-dimensional features. Figure 2.4 illustrates a CNN and its building blocks.

For 2D data, the convolutional layers comprise d learnable filters of a specific size, typically smaller than the input dimensions. These filters, which are of the form $D \times K \times K$, slide over the input data with a step size or stride s, computing the dot product of the filter weights and the respective region of the input data. The output of each filter is a spatial feature map $\boldsymbol{F}$ comprising the filter activations.

The number of filters d determines the number of distinct features extracted from the input data. A stride s greater than one reduces the spatial dimensions of the output volume. Conversely, a padding p of the input data with zeros can preserve the spatial dimensions of the output volume. Typically, each convolutional layer is followed by a non-linear activation function, such as the ReLU function. Furthermore, the addition of normalization layers, such as batch normalization [36], or LayerNorm [37], can be added to the network.

A typical CNN architecture comprises multiple convolutional and pooling layers, which extract features from the input data. Lastly, one or multiple FC layers are used to make predictions based on the extracted features. The number of ANs in the FC layers must be selected based on the target task. Several state-of-the-art CNN architectures have been developed, enabling stable training and robust feature extraction capabilities. Prominent examples are AlexNet [38], VGG [39], GoogLeNet [40], ResNet [41], DenseNet [42], EfficientNet [43], RegNet [44], and ConvNext [45].

Furthermore, CNNs can be employed for image segmentation tasks, wherein the objective is to assign a class label to each pixel of an image. To achieve a pixel-level prediction, the output layer of the CNN must have the exact dimensions as the input image. This is typically accomplished by transposed convolutional layers or upsampling layers to obtain feature maps of the desired spatial dimensions. A widespread architecture for medical image segmentation tasks is the Unet architecture [46], which consists of symmetrically mirrored down- and upsampling CNNs, also referred to as encoder and decoder, respectively. Skip connections facilitate information flow between the encoder and decoder at intermediate processing stages.

2.3.2 Training Deep Learning Models

The core of training deep learning models is the optimization algorithm used to adjust the model parameters, namely the weights and biases of the network, to minimize a predefined loss function $\mathcal{L}$.
The optimization algorithm or optimizer is typically a gradient-based algorithm that computes the gradient of the loss function with respect to the models' parameters θ

$$\nabla_\theta \mathcal{L} = \left(\frac{\partial \mathcal{L}}{\partial \theta_1}, \ldots, \frac{\partial \mathcal{L}}{\partial \theta_n} \right). \tag{2.6}$$

The gradients $\nabla_\theta \mathcal{L}$ are used to update the parameters θ

$$\theta_{t+1} = \theta_t - \alpha \nabla_\theta \mathcal{L}. \tag{2.7}$$

The learning rate α represents a crucial hyperparameter, defining the step size for each parameter update. A high learning rate can result in overshooting the minimum of the loss function, while a low learning rate can result in slow convergence of the optimization algorithm. To compute the gradients of the loss function, the backpropagation algorithm is used [47]. The backpropagation algorithm computes the gradients of the loss function with respect to the parameters of the network by applying the chain rule of calculus. The optimization algorithm is typically executed for multiple epochs, wherein an epoch is defined as a single pass through the training data. Training is terminated when a specified stopping criterion is met, such as reaching a predefined number of epochs or a minimum or saturating loss value on the validation data. As fitting all the training data into the memory is often impossible, the training data is typically divided into mini-batches. The optimization algorithm then computes the gradients of the loss function for each mini-batch and updates the parameters θ based on the average gradients of the mini-batches. This process is referred to as stochastic gradient descent. Several optimization algorithms have been developed to enhance the optimization process. In most of these extensions, historical values of the gradients are employed to adjust the learning rate [48], or first and second-order moments of the gradients are used to adapt the learning rate [49, 50].

2.4 Generative Models

The objective of GMs is to learn the underlying data distribution of the training data and to generate new data samples following the training distribution. Exemplary GMs include

Generative Adversarial Networks (GANs) [51], normalizing flows [52], autoregressive models such as PixelRNN [53] and variational autoencoders [49] or DDPMs [54, 55]. GMs have a wide range of applications, including image or text generation [51, 49, 53], text-to-image synthesis [56], image-to-image translation [57], data augmentation [58], super-resolution [59], and unsupervised anomaly detection [60, 61, 62]. In the following, we introduce the fundamental concepts of autoencoders, variational autoencoders and diffusion models, based on [29] and [63]. While autoencoders are not generally considered GMs, they are introduced here as they form the basis for their variational counterparts.

2.4.1 Autoencoders

Autoencoders (AE) are a class of neural networks that follow an encoder-decoder architecture. The AE framework is illustrated in Figure 2.5.
An encoder is trained to map an input data sample $\boldsymbol{x} \in \mathbb{R}^{H \times W}$ with height H and width W to a latent representation $\boldsymbol{z} = f_\phi^{enc}(\boldsymbol{x}) \in \mathbb{R}^d$. Typically, the latent dimension d is chosen to be lower than the input dimensionality. The latent representation $\boldsymbol{z}$ is then passed to a decoder, which aims to reconstruct the input data sample $\hat{\boldsymbol{x}} = g_\theta^{dec}(\boldsymbol{z}) = g_\theta^{dec}(f_\phi^{enc}(\boldsymbol{x}))$. The function f_ϕ^{enc} represents the encoder with parameters ϕ, and the function g_θ^{dec} represents the decoder with parameters θ. During training, the parameters ϕ and θ are adjusted to minimize the difference between the input data sample $\boldsymbol{x}$ and the reconstruction $\hat{\boldsymbol{x}}$. This leads to the loss function $\mathcal{L}_{Rec}$:

$$\arg\min_{\phi,\theta} \mathcal{L}_{Rec}(\boldsymbol{x}, g_\theta^{dec}(f_\phi^{enc}(\boldsymbol{x}))). \tag{2.8}$$

Common choices for $\mathcal{L}_{Rec}$ are the MSE or mean absolute error MAE. The encoder and decoder are typically implemented as CNNs for grid-like data, such as images, using transposed convolutions [64] or upsampling layers in the decoder path. The latent representation $\boldsymbol{z}$ is typically chosen to be of lower dimensionality than the input data and is called the bottleneck of the AE. The AE learns a compressed representation of the input data that captures the most important information required for the reconstruction. Additional regularization strategies such as denoising objectives can prevent overfitting and improve the robustness of the latent representations [65, 66, 67, 68].

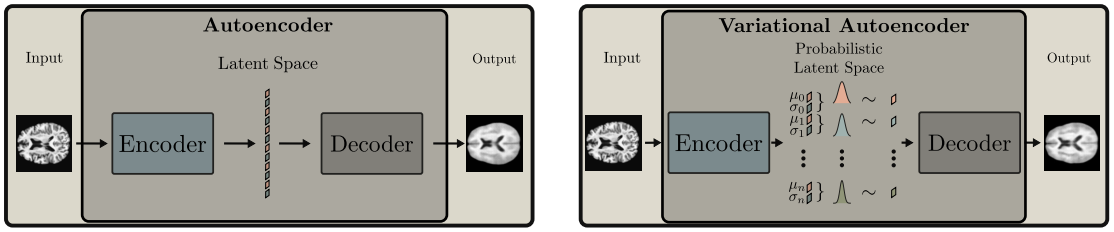


Fig. 2.5: Left: The autoencoder encodes the input data into a latent representation $\boldsymbol{z}$, and a decoder reconstructs the input data from the latent representation. Right: The variational autoencoder encodes the input data into the mean $\boldsymbol{\mu}_\phi(\boldsymbol{x})$ and the variance $\boldsymbol{\sigma}_\phi^2(\boldsymbol{x})$ of the approximate posterior distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$. The latent variable $\boldsymbol{z}$ is then sampled from this distribution, and a decoder reconstructs the input data from the sampled latent representation.

2.4.2 Variational Autoencoders

Variational AEs (VAE) [49] model the distribution of observed data $p(\boldsymbol{x})$ to generate new data samples. The observed data is often assumed to be generated by latent variables $\boldsymbol{z}$ that capture its underlying abstract features. The VAE framework is illustrated in Figure 2.5. Formally, the observed data $\boldsymbol{x}$ and its latent variables $\boldsymbol{z}$ can be described by the joint distribution $p(\boldsymbol{x}, \boldsymbol{z})$. The likelihood of the observed data $p(\boldsymbol{x})$ can then be expressed as:

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{z})\, dz \tag{2.9}$$

However, this marginalization is intractable for real-world data. Instead, the true posterior distribution $p(\boldsymbol{z}|\boldsymbol{x})$ is approximated with a variational distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$, parameterized by ϕ. This approximation allows optimizing a lower bound on the log-likelihood of the data, called the Evidence Lower Bound (ELBO). The derivation of the ELBO is based on Jensen's inequality and further reformulations, as shown in [49]. It can then be expressed as:

$$\log p(\boldsymbol{x}) \geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z})) \tag{2.10}$$

The first term, $\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]$ encourages accurate reconstructions of the input data given the sampled latent representation. Assuming a Gaussian distribution, this reconstruction term can be realized as the negative squared reconstruction error between the original input $\boldsymbol{x}$ and its reconstruction $\hat{\boldsymbol{x}}$. The second term, $D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z}))$, is the Kullback-Leibler divergence between the learned approximate posterior distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and the prior distribution $p(\boldsymbol{z})$. This term regularizes the learned latent distribution by encouraging it to stay close to a predefined prior, which is often chosen to be a standard normal distribution, $p(\boldsymbol{z}) = \mathcal{N}(0, I)$ as this enables an analytical solution.
Unlike standard AEs, which encode the input directly into a fixed latent vector, VAEs encode the input into two parameters: the mean $\boldsymbol{\mu}_\phi(\boldsymbol{x})$ and the variance $\boldsymbol{\sigma}^2_\phi(\boldsymbol{x})$ of the multivariate Gaussian distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ (the approximate posterior). The latent variable $\boldsymbol{z}$ is then sampled from this distribution as:

$$\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_\phi(\boldsymbol{x}), \boldsymbol{\sigma}^2_\phi(\boldsymbol{x}) \cdot \boldsymbol{I}) \tag{2.11}$$

In practice, a reparameterization trick is applied, enabling the backpropagation of the gradients. Instead of directly sampling $\boldsymbol{z}$, the latent variables are parametrized as:

$$\boldsymbol{z} = \boldsymbol{\mu}_\phi(\boldsymbol{x}) + \boldsymbol{\sigma}_\phi(\boldsymbol{x}) \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}), \tag{2.12}$$

where $\odot$ denotes the element-wise multiplication. This reparameterization expresses $\boldsymbol{z}$ as a differentiable function of $\boldsymbol{x}$ and auxiliary noise $\boldsymbol{\epsilon}$. The decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ is then used to reconstruct $\boldsymbol{x}$ from the latent variable $\boldsymbol{z}$. The final VAE objective can be expressed as:

$$\arg\min_{\phi,\theta} \left(\mathcal{L}_{Rec}(\boldsymbol{x}, p_\theta(\boldsymbol{x}|q_\phi(\boldsymbol{z}|\boldsymbol{x}))) + D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z}))\right) \tag{2.13}$$

Minimizing this objective allows the VAE to learn both a latent representation of the data and the generative process for sampling new data points from the learned latent space.
The VAE framework can be extended to Gaussian Mixture VAEs to model more complex data distributions [69].

2.4.3 Diffusion Models

Instead of encoding the input to a low-dimensional latent space, in DDPMs, the input is transformed into a Gaussian distribution through a sequence of noise-adding steps. The core idea is to learn how to reverse this noising process to generate realistic samples from pure noise. The training process of a DDPM is illustrated in Figure 2.7. The DDPM framework consists of two primary components: the forward process and the reverse process.

Forward Process

The forward process gradually adds Gaussian noise to an input data sample $\boldsymbol{x}_0 \in \mathbb{R}^{H \times W}$ with height H and width W over a series of T steps. At each time step t, noise is added to transform the original sample $\boldsymbol{x}_0$ into a noisy sample $\boldsymbol{x}_t$, increasing the amount of noise progressively.
Formally, the forward process can be described as:

$$\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}) \tag{2.14}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_t)$ represents the cumulative noise schedule, and β_t defines the noise level at each step. The noise schedule is predefined, and β_t is usually set such that $\boldsymbol{x_T} \sim \mathcal{N}(0, \boldsymbol{I})$ at the final step, meaning the sample becomes pure noise.

Reverse Process

In the reverse process, the goal is to recover the original noise-free sample $\boldsymbol{x}_0^{rec}$ from a noisy observation $\boldsymbol{x}_T$. The reverse process is defined as:

$$\boldsymbol{x}_0^{rec} \sim p_\theta(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t), \quad \text{where} \quad p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t))$$

Here, $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ and $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$ represent the mean and variance of the Gaussian distribution used to model the transition from $\boldsymbol{x}_t$ to $\boldsymbol{x}_{t-1}$. Often, only the mean $\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$ is learned during training, while the variance $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)$ is fixed to a predefined value based on the noise schedule:

$$\boldsymbol{\Sigma}(t) = \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t \boldsymbol{I}$$

For grid-like data such as images, the neural network used to model $\boldsymbol{\mu}_\theta$ is commonly implemented as a Unet. By iteratively applying these denoising steps, the model learns to generate new samples from noise. The forward and reverse process of a DDPM is illustrated in Figure 2.6.

Training Objective

Deriving the full training objective for DDPMs involves a series of reformulations and assumptions, which are beyond the scope of this dissertation. For a detailed derivation, we refer to [55, 63].
In essence, the training objective for DDPMs is to learn the reverse process by predicting the noise added to the sample during the forward process. At each time step t, the model

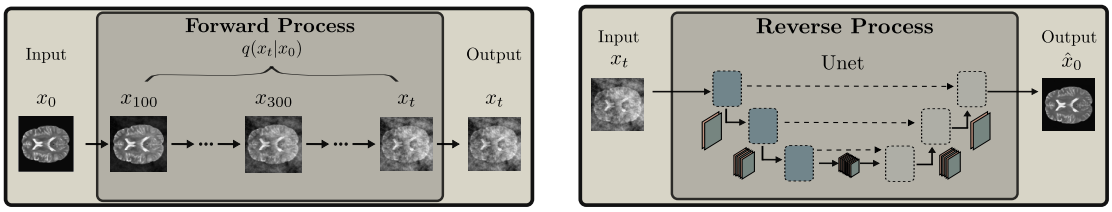


Fig. 2.6: Left: Forward process of a diffusion model. The input data sample $\boldsymbol{x}_0$ is progressively transformed to noised over a series of $t = 500$ steps. Right: Reverse process of a diffusion model. A Unet is employed to recover the original noise-free sample $\boldsymbol{x}_0$ from the noisy observation $\boldsymbol{x}_t$. Note that the Unet is conditioned by the time step t, which is omitted in this drawing for simplicity.

predicts the noise $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ that was added to $\boldsymbol{x}_t$. The loss function used to train DDPMs is the mean squared error between the true noise $\boldsymbol{\epsilon}$ and the predicted noise $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$:

$$\mathcal{L}_{\text{DM}} = \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\|^2 \tag{2.15}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$ is the noise sampled from a standard normal distribution, and $\boldsymbol{x}_t$ is the noisy version of $\boldsymbol{x}_0$ at time step t. During training, the time step t is typically sampled uniformly from $\{1, 2, ..., T\}$. Notably, rather than learning separate networks for each of the T transitions, a single network is trained to predict the transitions, with its predictions conditioned on the respective timestep t. Alternatively, instead of training the Unet to predict the noise that has been added, it can be trained to predict the denoised image directly. In this thesis, we adopt the latter approach, which is also reflected in the following explanation of the sampling process.

Sampling

The sampling process of a DDPM is illustrated in Figure 2.7. Once the model has been trained, the forward and reverse processes are used to generate new data by gradually denoising pure noise $\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I})$. At each step, noise is removed from $\boldsymbol{x}_t$ while reintroducing a smaller amount of noise to $\boldsymbol{x}_{t-1}$ until $\boldsymbol{x}_0$ is reached. In contrast to image synthesis, reconstruction-based UAD requires reconstructing a given image rather than generating entirely new samples. To achieve this, the input image is often partially

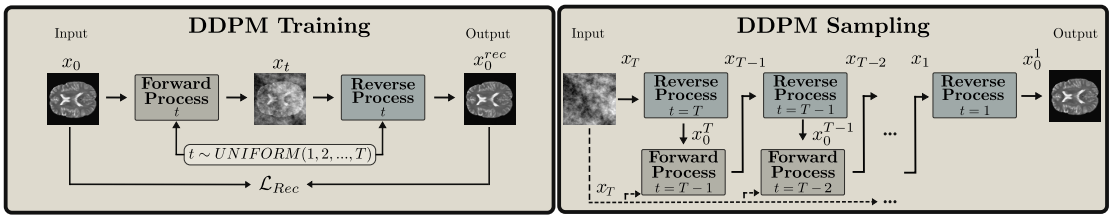


Fig. 2.7: Left: Training process. The forward and reverse process are applied to learn the transition from the original sample $\boldsymbol{x}_0$ to the noisy sample $\boldsymbol{x}_t$ and back. Right: Sampling process. The reverse and forward process are applied iteratively to generate new samples, starting with pure noise $\boldsymbol{x}_T$. The output of the reverse process at time step t is denoted by $\boldsymbol{x}_0^t$.

replaced with noise, and the sampling process begins from a partially noised input $\boldsymbol{x}_t$ instead of pure noise $\boldsymbol{x}_T$. The reverse process is a computationally expensive operation as it involves T, or t iterations for synthesis and reconstruction, respectively. To accelerate this sampling process, various efficient sampling techniques have been proposed [70, 71]. Alternatively, when starting from a partially noised image $\boldsymbol{x}_t$, the reverse process can be performed in a single step, which has been shown to enhance both anomaly detection performance and computational efficiency [72].

2.5 Anomaly Detection

Anomaly detection (AD) is a machine learning task that aims to detect and localize anomalies in data. AD can be applied to a wide range of problems [73], including fraud detection [74, 75, 76], genetics [77, 78], and industrial quality control [79, 80, 81, 82]. This concept is of particular value in medical imaging, as anomalies frequently indicate the presence of pathologies. In this section, we present a formal definition of AD based on the work of [83] and introduce AD approaches in the domain of brain MRI.

Anomalies can be defined as deviations from a baseline normality concept, which is determined by the available data and the context of the task. Importantly, anomalies do not necessarily represent pathological changes but can manifest, e.g., as artifacts introduced during image acquisition.
Formally, given a data space $\mathcal{X} \subseteq \mathbb{R}^D$, the concept of normality can be defined as distribution $\mathbb{D}^+$ on the data space $\mathcal{X}$. If a data sample $x \in \mathcal{X}$, or a set of data samples, does not lie in the distribution $\mathbb{D}^+$, the sample is considered an anomaly. Given the probability density function $d^+(x)$ of $\mathbb{D}^+$, the set of anomalies can then be defined as

$$\mathcal{A} = \{x \in \mathcal{X} \mid d^+(x) < \epsilon\}, \tag{2.16}$$

where ϵ is a threshold that defines the boundary between normal and anomalous data samples.
The goal of an AD model is then to score potential anomalies $\tilde{x} \in \mathcal{X}$ based on the learned normal data distribution. Therefore, the objective of AD models is to estimate the low-density regions in the data space $\mathcal{X}$ under $\mathbb{D}^+$. Formally, this can be expressed as density level set estimation whereby the objective is to find the density level set C_α of the normal data distribution $\mathbb{D}^+$ that contains a fraction α of the normal data samples. The density level set C_α can be defined as

$$C_\alpha = \{x \in \mathcal{X} \mid d^+(x) \geq \epsilon_\alpha\}, \tag{2.17}$$

where ϵ_α is the threshold that defines the boundary of the density level set C_α. Given the concentration assumption [84] holds, some level α exists such that the density level set C_α exists and can be bounded. For such a given set C_α, a threshold anomaly detector $c_\alpha : \mathcal{X} \to \{0, 1\}$ could be defined as

$$c_\alpha(x) = \begin{cases} 1 & \text{if } x \in C_\alpha, \\ -1 & \text{otherwise.} \end{cases} \tag{2.18}$$

The normal data distribution $\mathbb{D}^+$ is often unknown and must be estimated from the available data. In medical imaging, $\mathbb{D}^+$ is often defined by the data obtained from

healthy patients. Anomalies are then considered to be deviations from this distribution. The majority of AD approaches are performed in an unsupervised setting, where only unlabeled data samples $\{(x_i)\}_{i=1}^N \in \mathcal{X}$ are available. Although the data distribution $\mathbb{D}^+$ may be contaminated with anomalies, resulting in a mixture with contamination rate $\eta > 0$, unsupervised methods often assume no anomalies are present ($\eta = 0$).
Additionally, AD can be applied in semi supervised settings, wherein a subset $\{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^M \in \mathcal{X} \times \mathcal{Y}$ of the data samples $\{(x_i)\}_{i=1}^N \in \mathcal{X}$, contains additional information $\mathcal{Y}$ that differentiates anomalies ($\tilde{y}_i = 1$) from normal data samples ($\tilde{y}_i = 0$). This labeling facilitates the calibration of models or the detection of specific anomalies. Less commonly, AD is applied in the completely supervised setting, which entails a binary classification task necessitating labeled data $(\tilde{x}_i, \tilde{y}_i)_{i=1}^N \in \mathcal{X} \times \mathcal{Y}$. Since anomalies are often rare and not well-defined in advance, unsupervised AD or UAD is the most common approach.

2.5.1 Approaches to Unsupervised Anomaly Detection

As previously outlined, an effective methodology for UAD entails estimating the density function, which serves as a foundation for identifying anomalies.
Formally, given a new data sample x', the anomaly score $s(x')$ is given by the density function $\hat{d}(x'; x_1 \dots x_n)$. By applying $\hat{d}(x'; x_1 \dots x_n)$ in Equation 2.17, we can estimate the density level set C_α. This level set can then be used in Equation 2.18 to detect anomalies. Density estimation can be achieved by using parametric density estimation methods, such as Gaussian Mixture Models [85], Kernel Density Estimation [86, 87], or VAEs [49]. However, particularly when pixel-wise predictions are required, the high-dimensional data space can make density estimation challenging [88]. Moreover, estimating the full density function is not always necessary. In some cases, identifying the low-density regions within the data space $\mathcal{X}$ is sufficient for UAD [89, 90]. Therefore, an alternative approach is to use one-class classification methods [91, 89], which learn a decision boundary around the data distribution $\mathbb{D}^+$ and classify data samples that lie outside the decision boundary as anomalies. Formally, one-class classifiers aim to find a decision function f^{OC} such that

$$f^{OC}(x') = \begin{cases} 1 & \text{if } x' \in \mathbb{D}^+, \\ -1 & \text{otherwise.} \end{cases} \tag{2.19}$$

While this approach only provides a binary decision for each data sample, a continuous anomaly score $s(x')$ can be derived by measuring the distance to the decision boundary [92]. Typically, one-class classification is applied in some feature or kernel space, where the data samples are mapped to a higher-dimensional feature space, and a decision boundary is learned in this space. Common examples of one-class classification methods applied in a feature space are Support Vector Data Description (SVDD) [90] or one-class Support Vector Machines (SVM) [89]. Learning the feature space using a deep learning model is also possible, leading to deep SVDD [93] or deep one-class SVMs [94].
Another approach to UAD involves reconstruction-based methods. In this context, a GM $f^{GM} : \mathcal{X} \to \mathcal{X}$ is trained to minimize the reconstruction error L_{rec} between input data samples $x_1, \dots, x_n \in \mathcal{D}^+$ and their reconstructions $\hat{x}_1 = f^{GM}(x_1), \dots, \hat{x}_n = f^{GM}(x_n) \in \tilde{\mathcal{D}}^+$. When applied to new data, reconstruction errors can then indicate anomalies.
Formally, the anomaly score for a new data sample x' is given by the reconstruction error $s(x', \hat{x}') = L_{rec}(x', \hat{x}')$. A new sample x' is considered an anomaly if the reconstruction

error exceeds a threshold ϵ. Thus, the criterion is

$$c(x') = \begin{cases} 1 & \text{if } s(x', \hat{x}) \leq \epsilon, \\ -1 & \text{otherwise.} \end{cases} \tag{2.20}$$

The threshold ϵ defines the boundary between data samples that may be considered normal or anomalous. In this approach, rather than estimating the density $\hat{d}(x')$ of individual samples, the GM f^{GM} is employed to transform any sample $x' \in \mathcal{X}$ to a reconstruction $\hat{x} = f^{GM}(x')$ in the data space $\tilde{\mathcal{D}}^+$ and the anomaly score $s(x)$ is obtained directly from the reconstruction error $L_{rec}(x, f^{GM}(x))$. Common GMs for reconstruction-based UAD are AEs, VAEs [62] or GANs [60]. While certain GMs, such as VAEs, are capable of approximating the density function $\hat{d}(x; x_1 \dots x_n)$, empirical evidence suggests that the reconstruction error L_{rec} is often a more reliable indicator for anomalies than the estimated density function [95].

2.5.2 Unsupervised Anomaly Detection in Brain MRI

Considering the application of anomaly detection in brain MRI, the high dimensionality of the data space makes direct density estimation challenging. Consequently, MRI scans are typically reduced to a lower-dimensional feature space prior to the application of anomaly detection methods. Throughout this thesis, we refer to this approach as feature-based approach. The most common strategy to obtain meaningful features from brain MRI is to use the latent representation of AEs. In this feature space, primarily one-class classification methods have been applied to detect anomalies [96, 97]. Furthermore, throughout the progress of this thesis, approaches that rely on density estimation [98, 99] or discrepancies between latent representations have been proposed [82, 100, 101].
Reconstruction-based anomaly detection has been shown to be particularly effective for high-dimensional brain MRI scans. In contrast to the feature-based approach, reconstruction-based anomaly detection directly employs the reconstruction error of the GM as the anomaly score. Thereby, anomalies can be detected on a pixel-wise or voxel-wise level, allowing for a fine-grained localization of anomalies in the brain MRI scans. Several studies have explored reconstruction-based anomaly detection in brain MRI using AEs [102, 103, 104, 105, 106, 62], VAEs [107, 108, 109, 61, 110] and GANs [111, 112, 113, 114, 4]. More recently, denoising AEs have been applied in this context [115] and DDPMs emerged as a powerful class of GMs for UAD [116, 117]. Additionally, self supervised learning with synthetic anomalies has been proposed as another approach to UAD in brain MRI [118, 119, 120, 121]. A detailed discussion of the related work in the field of UAD in brain MRI is provided in Chapter 5.
In this thesis, we focus on applying reconstruction-based anomaly detection to brain MRI. Specifically, we investigate using VAEs and DDPMs as GMs and propose novel adaptations that integrate additional contextual information into the reconstruction process and anomaly scoring mechanism.

3 Methods and Experiments

In this chapter, we present the details of our experimental setup. We begin with an overview of the used data sets, followed by a description of the applied pre-processing and post-processing steps. We then introduce the metrics and baselines that are evaluated. Finally, we present the motivation and the details of our proposed approaches.

3.1 Data Sets

For training, data sets consisting of only healthy samples are used, while evaluation is performed on data sets containing various pathologies along with their corresponding annotations. For training, we utilize the following data sets:

MixedNormals Data Set The MixedNormals (MN) data set is a proprietary data set provided by Jung Diagnostics GmbH and consists of 1971 T1-weighted brain MRI scans. At the time of data acquisition, none of the patients had a documented history of neurological or psychiatric conditions. For all patients, a local radiologist conducted a visual assessment and confirmed that all volumetric images showed no abnormalities beyond age-related variations. The MN data set comprises scans obtained with different acquisition devices from distinct vendors, encompassing varying acquisition parameters. We separate a test set of 10 % of the data set for evaluation. The remaining data is used for training and validation.

IXI Data Set The Information eXtraction from Images (IXI) data set [122] is publicly accessible and comprises 562 healthy brain MRI scans. We separate a healthy test set consisting of 160 samples. The remaining data is partitioned into five training sets (N=358) and validation sets (N=44) for cross-validation. The scans are acquired by three different scanner models with varying acquisition parameters and contrast settings, including T1-weighted and T2-weighted scans.

For evaluation, we use the following data sets:

BRATS Data Set The BRATS data set is a publicly accessible data set for the multimodal BRAin Tumour Segmentation (BRATS) Challenge [9, 123, 28]. The data set is updated on an annual basis, resulting in the release of multiple versions. The 2019 Version comprises 335 annotated brain MRI scans, whereas the 2021 Version comprises 1250 annotated brain MRI scans. For all data sets, voxel-level annotation masks are provided by radiologists as categorical masks. Two distinct types of tumors are present in the BRATS data sets, namely glioblastoma and lower-grade glioma. For each subject, T1, T1 contrast-enhanced, T2 and FLAIR weighted scans are available.

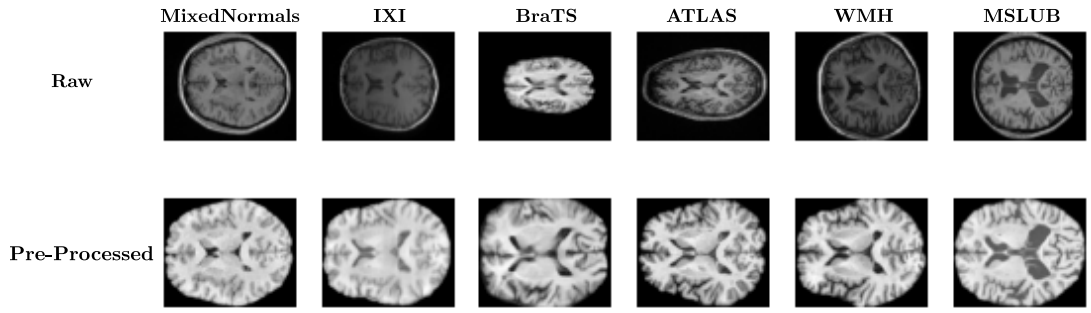


Fig. 3.1: Overview of the data sets used in the experiments. Top: Center slices of exemplary raw MRI scans before pre-processing, resampled to a uniform voxel grid. Bottom: Center slices of the corresponding pre-processed MRI scans.

MSLUB Data Set The Multiple Sclerosis data set from the University Hospital of Ljubljana (MSLUB) [124] is publicly accessible and comprises 30 multimodal scans of subjects diagnosed with Multiple Sclerosis (MS) and the corresponding annotations. T1, T2 and FLAIR weighted scans are available for each subject.

ATLAS Data Sets The ATLAS data set [125] is a publicly accessible data set for the detection and segmentation of stroke lesions. Two versions of the data set are available, both containing T1-weighted MRI scans. The first version (v1) comprises 304 annotated samples with stroke lesions, while the second version (v2) consists of 655 annotated samples. The stroke lesions are annotated by experts and binary segmentation masks are provided. The samples predominantly exhibit embolic strokes in the cortical and subcortical regions of the left and right hemispheres.

WMH Data Set The White Matter Hyperintensity (WMH) data set [126] encompasses 60 MRI scans from patients with WMH. Segmentation masks are derived from consensus agreements between two expert radiologists. T1 and FLAIR weighted scans are available for each subject.

3.1.1 Pre-Processing

We apply the following pre-processing steps to bring the data sets into a common format. We pre-process all scans by resampling them to an isotropic resolution of $1 \times 1 \times 1$ mm and registering them to the SRI24-Atlas template [127]. The brain region is extracted using a Unet [128], after which the images are cropped to the size of the brain region. To correct for intensity inhomogeneities, N4 bias field correction [129] is applied. The images are then cropped or padded to a fixed size of $192 \times 192 \times 160$ voxels. Intensity values are normalized to the range $[0, 1]$, based on the 99% percentiles of the intensity distribution. Next, the dimensions of the images are reduced by a factor of two to $96 \times 96 \times 80$ voxels. Finally, 15 top and bottom slices are cropped, resulting in a final image size of $96 \times 96 \times 50$ voxels.
Pre-processing is performed using the ANTS [130] and TorchIO [131] libraries. Figure

3.1 shows a visualization of the pre-processed images. The pre-processing steps are applied to all data sets used in the experiments except for our initial study [20] where we follow the pre-processing steps of [62] and further reduce the dimension of the images to $64 \times 64 \times 64$.

3.1.2 Post-Processing

We apply the following post-processing steps to the anomaly maps to evaluate the proposed approaches and the baselines. We apply a median filter with a kernel size of $K = 5 \times 5 \times 5$ to smooth the residual map. Following this, we perform brain mask erosion for three iterations, primarily aimed at filtering out residuals caused by poor reconstructions at sharp edges near the brain mask [62]. The residual map is then binarized using a greedy threshold search [109]. After binarization, connected component filtering is applied to remove areas with fewer than seven voxels.

3.2 Evaluation

We conduct various experiments and evaluate different metrics to assess our proposed approaches.

Reconstruction Metrics

To assess the reconstruction quality, we employ the validation or test set of the healthy IXI or MN data sets and calculate similarity metrics between the input and the reconstruction. We consider the Structural Similarity Index Measure (SSIM) [132], the Peak Signal to Noise Ratio (PSNR) and the Learned Perceptual Image Patch Similarity (LPIPS) [133] as metrics to evaluate the reconstruction quality. Moreover, we consider the reconstruction error ($l1$-error). Given that the GMs are trained to reconstruct healthy anatomy, it is essential to consider the $l1$-error of healthy and unhealthy anatomy separately. Therefore, for the unhealthy evaluation data sets, we calculate the $l1$-error for healthy and unhealthy anatomy, as indicated by the annotation masks and introduce an $l1$-ratio as follows:

$$l1\text{-ratio} = \frac{l1_{unhealthy}}{l1_{healthy}}.$$

A higher value for the $l1$-ratio indicates that the model successfully reconstructs the healthy anatomy without replicating the unhealthy parts of the input. A lower value indicates that the model fails to reconstruct the healthy anatomy or replicates the unhealthy parts of the input.

Segmentation Metrics

To assess the segmentation performance, we report the Dice score [134, 135] (DICE). The Dice score is a widely used metric for segmentation tasks, providing a measure of the overlap between the predicted and the ground truth segmentation. The Dice score requires binary masks as input, where the predicted anomaly map is binarized using a threshold. We consider two versions of the Dice score. First, an unhealthy validation set is employed to identify the optimal threshold, which is subsequently applied to the test set. This version is designated as DICE. Secondly, we calculate an optimal Dice score by

searching for the threshold that maximizes the Dice score for the test data. This version is designated as ⌈DICE⌉. The thresholds are estimated by a greedy search, following the method described in [109]. Since the threshold search is performed on the validation or test data, the resulting Dice scores represent the upper bounds for the segmentation performance.
To obtain a threshold-independent metric, we calculate and report the Area Under Precision-Recall Curve (AUPRC). The AUPRC quantifies the precision-recall trade-off and is particularly useful for imbalanced data sets. The AUPRC is calculated by sorting the anomaly scores in descending order and calculating the precision and recall for each threshold. The AUPRC is then obtained by integrating the precision-recall curve.

Statistical Testing

To assess the statistical significance of performance differences between models, we employ a permutation test from the MLXtend library [136]. Specifically, we conduct 10,000 permutations at a significance level of $\alpha = 5\%$. In each permutation, the mean difference between the scores of the two models is computed. The p-value is then determined by counting the times the permuted mean differences are greater than or equal to the observed sample difference, normalized by the total number of permutations.

3.3 Baselines

We evaluate our proposed approaches in comparison with established baselines for UAD in brain MRI. These include feature-modeling methods, reconstruction-based approaches, and training strategies that utilize synthetic anomalies.

Feature-modeling Methods

We evaluate the following feature-modeling methods that rely on feature discrepancies for anomaly scoring:
Feature Autoencoder (FAE) [101]: In this method, the input image is first transformed into a feature space using a pre-trained CNN. Then, an AE is employed to reconstruct the extracted features at multiple resolutions. The features and their corresponding reconstructions are then resized to match the original image size, and the SSIM between the pairs is used as the anomaly score.
Reverse Distillation (RD) [82]: Reverse Distillation (RD) combines an AE architecture with a student-teacher knowledge distillation approach. Given a teacher encoder pre-trained on mixed data, a student decoder is trained to reconstruct the representations produced by the teacher encoder for healthy samples. The upsampled cosine distance between the feature maps of the teacher encoder and student decoder is used as the anomaly score.
Encoder Decoder Consistency (EDC) [100]: The EDC method utilizes the contrast between encoder and decoder features in AEs. Unlike conventional feature-based methods that use frozen pre-trained encoders from natural image data sets, EDC optimizes the encoder and decoder end-to-end within the medical imaging domain, integrating contrastive learning. Anomaly scores are computed using the upsampled cosine similarity between encoder and decoder feature representations.

Reconstruction-based Approaches

We evaluate the following reconstruction-based approaches:
Autoencoder (AE) [62]: The AE is a simple baseline that uses a convolutional AE to reconstruct the input image. The anomaly score is the pixel-wise $l1$-error between the input and the reconstruction.
Variational Autoencoder (VAE) [62]: The VAE is an extension of the AE that learns a probabilistic latent space representation from which the input data is reconstructed. The anomaly score is the pixel-wise $l1$-error between the input and the reconstruction.
Sequential VAE (SVAE) [137] The SVAE is a variant of the VAE that aims to capture inter-slice dependencies within an MRI volume. SVAEs consist of a 2D encoder that extracts features for each input slice. The features are then processed as a sequence by a transformer network to model and capture the dependencies across slices. After the sequence processing, the features are transformed into the image space by a 2D decoder. Thereby, 3D information is captured in the latent space. The anomaly score is the pixel-wise $l1$-error between the input and the reconstruction.
Denoising Autoencoder (DAE) [115]: The DAE is a variant of the AE that is trained to reconstruct the input from corrupted versions of the input. In this implementation, a Unet is used as the AE architecture and upscaled Gaussian noise is added to the input. The anomaly score is the pixel-wise $l1$-error between the input and the reconstruction. Notably, during evaluation, no noise is applied to the input.
Reverse Anomalies (RA) [138]: The RA method uses a soft intro VAE [139] with a multi-scale embedding loss to compare input and reconstruction features at multiple encoder stages. This combination aims to improve the generation of healthy anatomy and to enhance the coherence between input and reconstruction. The $l1$-error is then combined with the upsampled pixel-wise LPIPS for anomaly scoring.
PHANES [140]: PHANES is a hybrid approach that integrates a GAN with RA for inpainting tasks. The method involves generating binary masks given the residual map of the RA method and using them for GAN-based inpainting. The $l1$-error is then combined with the upsampled pixel-wise LPIPS for anomaly scoring.
AnoDDPM [116]: AnoDDPM is a diffusion-based approach where the Gaussian noise added during the forward process is replaced by structured Simplex noise. Furthermore, instead of starting the sampling from pure noise, the sampling starts with a partially noised image. Notably, we use the denoising Unet to generate a reconstruction from a given image $\boldsymbol{x}_{t_{test}}$ in a single step, setting $t_{test} = \frac{T}{2} = 500$. The anomaly score is $l1$-error between the input and the reconstruction.

Synthetic Anomalies

We evaluate the following methods that utilize synthetic anomalies for training:
Foreign Patch Interpolation (FPI) [120]: In FPI, foreign image patches are blended into an image to generate artificial anomalies. The patch regions are extracted from two independent samples at the same location and replaced with an interpolation between both patches. A Unet-like architecture is trained to predict the location and interpolation factor of the patch. The predictions of the Unet are then used as the anomaly score.
DRAEM-Net [141]: DRAEM-Net employs a dual-network architecture comprising a generator and a segmentation network. The generator is trained to remove synthetic anomalies, providing a pseudo-healthy reconstruction. The segmentation network is then

used to segment the synthetic anomalies, given the concatenation of abnormal input and pseudo-healthy reconstruction. The predictions of the Unet are then used as the anomaly score.

Learning-free Baseline

We additionally evaluate a learning-free baseline that does not require the training of deep learning models.
Intensity-based Thresholding (Thresh) [142]: This method is based on simple thresholding. Histogram Equalization is applied to the input image, and a threshold is determined by an optimization process on the validation set. The anomaly score is then directly derived from the binarized image.

3.4 Proposed Approaches

In this section, we introduce our proposed approaches and outline their motivation. For all approaches, a detailed description of the methodology is provided in the corresponding publications, attached in Chapter 8.

3.4.1 3D VAEs with Spatial Erasing

This approach is based on our study *3-Dimensional Deep Learning with Spatial Erasing for Unsupervised Anomaly Segmentation in Brain MRI* [20], attached in Chapter 8.1.
VAEs are predominantly used GMs for reconstruction-based UAD in brain MRI. However, while the strong regularization of VAEs prevents the replication of unhealthy structures, their reconstructions often lack fine anatomical detail and are blurry [143, 62, 144]. While several methods have been proposed to enhance the quality of VAE reconstructions and anomaly scoring [108, 109, 61], most of the proposed methods neglect the 3D information of MRI data. To address this shortcoming, we investigate using 3D VAEs to incorporate additional context in the form of volumetric information. We hypothesize that 3D spatial information will improve UAD performance due to increased reconstruction fidelity. Given the larger parameter space in 3D models, we also introduce spatial erasing to mitigate the risk of overfitting and to enforce the utilization of 3D context.

Approach

Our overall approach [20] is shown in Figure 3.2. We extend the baseline 2D VAE by applying 3D operations, such as 3D convolutions, 3D normalization or 3D pooling. Furthermore, we extend existing 2D erasing techniques from cutout [145] and context autoencoders [68, 109] to 3D MRI scans, proposing several erasing strategies:
Single Patch/Cube Erasing: One patch or cube is randomly erased, ranging from 1% to 25% pixels or voxels of the input image.
Multiple-Patch/Cube Erasing: Multiple randomly positioned patches or cubes are erased, totaling 1% to 25% of the input size.
Half-Slice/Volume Erasing: One side of the brain is randomly erased in a 2D slice or across multiple consecutive 3D slices.

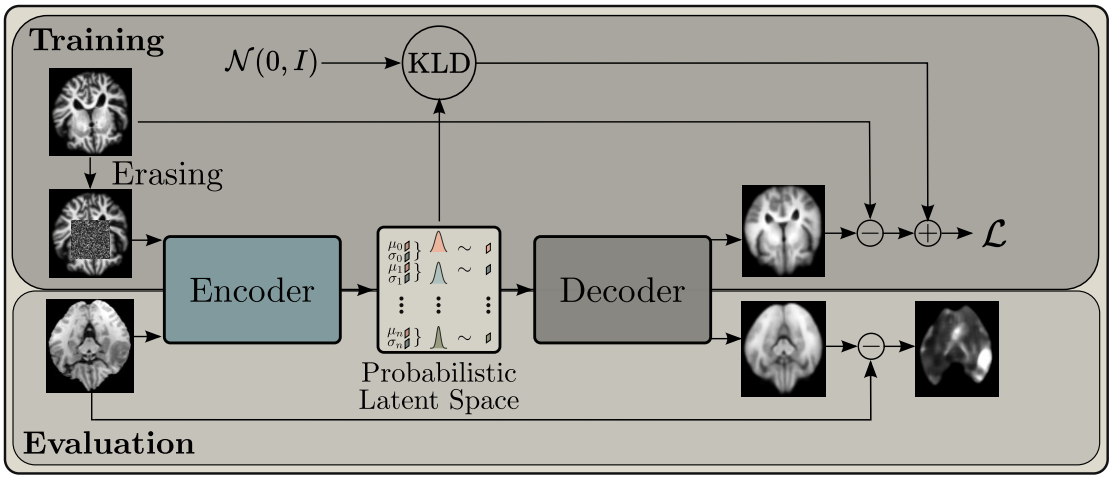


Fig. 3.2: Our proposed erasing approach, adapted from [20]. During training, the input image is erased by a patch or cube, and the VAE is trained to reconstruct the original image. At test time, no erasing is applied, and the residual map between input and reconstruction is used as the anomaly score.

During training, we apply erasing strategies with a binary mask $\boldsymbol{M}_e$, where ones indicate the erased region and zeros indicate the preserved background creating an erased input $\tilde{\boldsymbol{x}} = \boldsymbol{x} \odot \neg \boldsymbol{M}_e$. The loss function in Equation 2.13 then becomes:

$$\arg\min_{\phi,\theta} \left(\mathcal{L}_{Rec}(\boldsymbol{x}, p_\theta(\boldsymbol{x}|q_\phi(\boldsymbol{z}|\tilde{\boldsymbol{x}}))) + D_{KL}(q_\phi(\boldsymbol{z}|\tilde{\boldsymbol{x}}) \| p(\boldsymbol{z})) \right)$$

Optionally, erased regions are replaced with Gaussian noise instead of zeros. At test time, we use the residual map based on the l_1-error between the input and reconstruction to calculate anomaly scores.

3.4.2 Patched Diffusion Models

This approach is based on our study *Patched Diffusion Models for Unsupervised Anomaly Detection in Brain MRI* [21], attached in Chapter 8.2.

The blurriness of VAE reconstructions can be largely attributed to the loss of spatial information within the fully connected, dense latent space. Studies have shown that spatial latent representations (i.e., two-dimensional feature maps) can substantially enhance reconstruction quality [62, 146]. However, the dense latent spaces in classical VAEs present an important bottleneck and thus regularization that prevents the model from simply copying the input [62]. Therefore, the utilization of spatial latent spaces or skip connections necessitates the incorporation of additional regularization.

DDPMs address this issue by introducing noise into the input image and reconstructing it with a Unet-based architecture, demonstrating robust outcomes in image synthesis [55] and UAD [116]. Nevertheless, the accurate reconstruction of fine-grained details of the brain from a noisy image remains a challenge. In particular, besides the loss of structural information, applying noise to the entire input image can result in the loss of contrast and intensity information. This loss of information can lead to inconsistencies in the reconstructed image.

To address these limitations, we propose patched Denoising Diffusion Probabilistic

Models (pDDPM), which aim to use the global context of the clean target image in the denoising process. Specifically, the forward process is applied only to a predefined patch within the input image, while the remaining image is unaltered. The backward process is then applied to the entire image to recover the noised patch. We hypothesize that using the unaltered surrounding context during denoising improves reconstruction quality and coherence of input and reconstructions, enabling more accurate anomaly scoring.

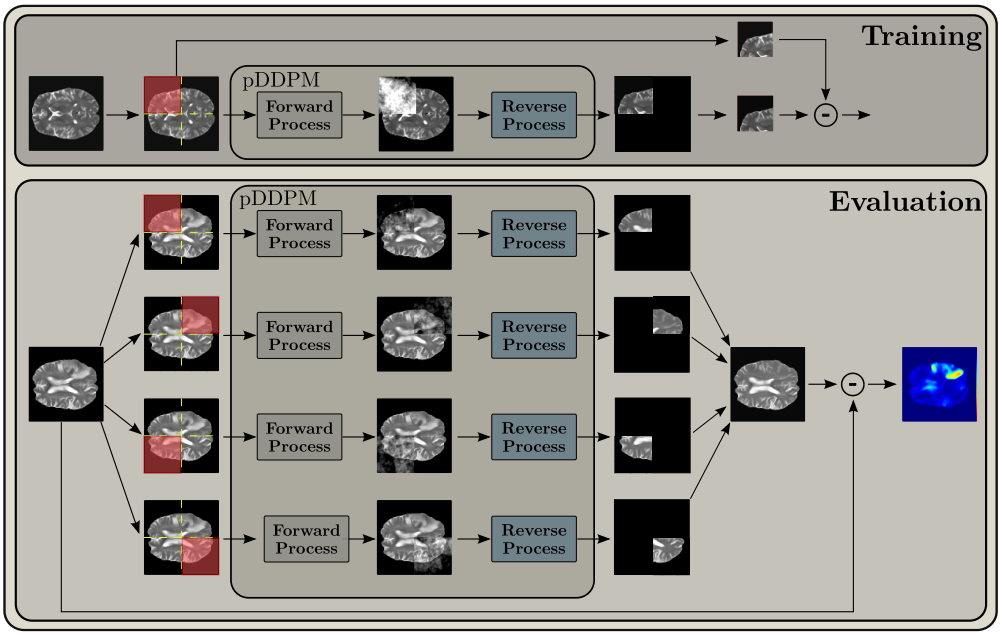


Fig. 3.3: Overview of our pDDPM, adapted from [21]. During training (top), a single patch is sampled from the input image, subjected to noise in the forward process, and denoised in the backward process. During evaluation (bottom), multiple patches are subjected to noise and the reconstructed patches are stitched together to form a complete reconstruction. The $l1$-error is then computed as the anomaly map.

Approach

The pDDPM, illustrated in Figure 3.3, builds upon the AnoDDPM [116], using simplex noise instead of Gaussian noise.
Given an input image $\boldsymbol{x} \in \mathbb{R}^{H \times W}$ with height H and width W, we partition $\boldsymbol{x}$ into K patch regions $\boldsymbol{p}_k \in \mathbb{R}^{h \times w}$ with $h < H$ and $w < W$. During training, patches are sampled either at random positions or from a fixed grid. For the grid, K patch regions are evenly spaced across $\boldsymbol{x}$, with the number of possible patches calculated as $K = \lceil \frac{W-w}{w} \rceil + \lceil \frac{H-h}{h} \rceil + 2$, where $\lceil . \rceil$ denotes the ceiling function. We then uniformly sample an index k from this grid.
Given a selected patch region $\boldsymbol{p}_k$, we apply the forward process to introduce noise to this region as follows: starting with the original image $\boldsymbol{x}$ and a noise level t, we obtain the partially noised image $\boldsymbol{x}_t$ following Equation 2.14. Using a binary mask $\boldsymbol{M}_p \in \mathbb{R}^{H \times W}$, where ones indicate the patch region and zeros indicate the background, we define the

partly noised image:

$$\tilde{\boldsymbol{x}}_t = \boldsymbol{x}_t \odot \boldsymbol{M}_p + \boldsymbol{x} \odot \neg \boldsymbol{M}_p.$$

Here, $\odot$ represents element-wise multiplication. The partly noised image $\tilde{\boldsymbol{x}}_{\boldsymbol{t}}$ is then passed through the denoising Unet, which predicts the original content within the patch, leading to $\tilde{\boldsymbol{x}}^{\text{rec}}$. To concentrate the loss on the patch-specific denoising task, we optionally adapt the reconstruction loss $\mathcal{L}_{\text{rec}}$ to a patch-wise loss $\mathcal{L}_p$ focused only on the patch region:

$$\mathcal{L}_p = \left| (\boldsymbol{x} - \tilde{\boldsymbol{x}}^{\text{rec}}) \odot \boldsymbol{M}_p \right|.$$

At test time, we sequentially apply the forward and backward processes to all K patches in the input image using a sliding window approach, allowing each patch to be estimated with spatial context from surrounding patches. The whole image is then reconstructed by stitching together the individual denoised patches, averaging in overlapping regions to smooth transitions. For each patch, we use the denoising Unet to generate a reconstruction from $\boldsymbol{x}_{t_{test}}$ in a single step, as this has shown to improve performance while reducing processing time. Initially, we evaluate the model using a single noise level, set to $t_{test} = \frac{T}{2} = 500$, and subsequently conduct ablation studies to assess the impact of different noise levels. Finally, the residual map between the input and the reconstructed image serves as the anomaly score.

3.4.3 Context-Conditioned Diffusion Models

This approach is based on our study *Guided Reconstruction with Conditioned Diffusion Models for Unsupervised Anomaly Detection in Brain MRIs* [22], attached in Chapter 8.3.

As demonstrated in our study [21], the patching strategy in pDDPMs can enhance the reconstruction quality. However, the patching approach increases computational complexity and processing time, as each image is processed in multiple iterations. Furthermore, the patching technique may result in the introduction of artifacts in regions of overlap and necessitates tuning an additional hyperparameter, namely the patch size. To address these limitations, we propose context-conditioned DDPMs (cDDPMs). Our approach introduces a conditioning mechanism that provides the denoising process with additional context from the input image for guidance. An additional encoder network provides an abstract representation, or context vector, to condition the denoising process. Consequently, image context can be integrated into the denoising process without the need for a costly patching strategy.

We hypothesize that this conditioning mechanism improves alignment in local intensity distributions between input and reconstruction, resulting in enhanced reconstruction quality and segmentation performance. Additionally, the conditioning mechanism, which is adaptive to the input image, could enhance the model's generalization and domain adaptation capabilities.

Approach

The cDDPM architecture is illustrated in Figure 3.4. Given an input image $\boldsymbol{x}$, we apply the DDPM forward process from Equation 2.14 to obtain a noised version $\boldsymbol{x}_t$. In addition, we derive a context vector $\boldsymbol{c}$ by encoding $\boldsymbol{x}$ through a ResNet-based encoder network, $\boldsymbol{c} = F_{\text{enc}}(\boldsymbol{x})$, where $\boldsymbol{c} \in \mathbb{R}^d$ and d is the dimension of the context vector. We

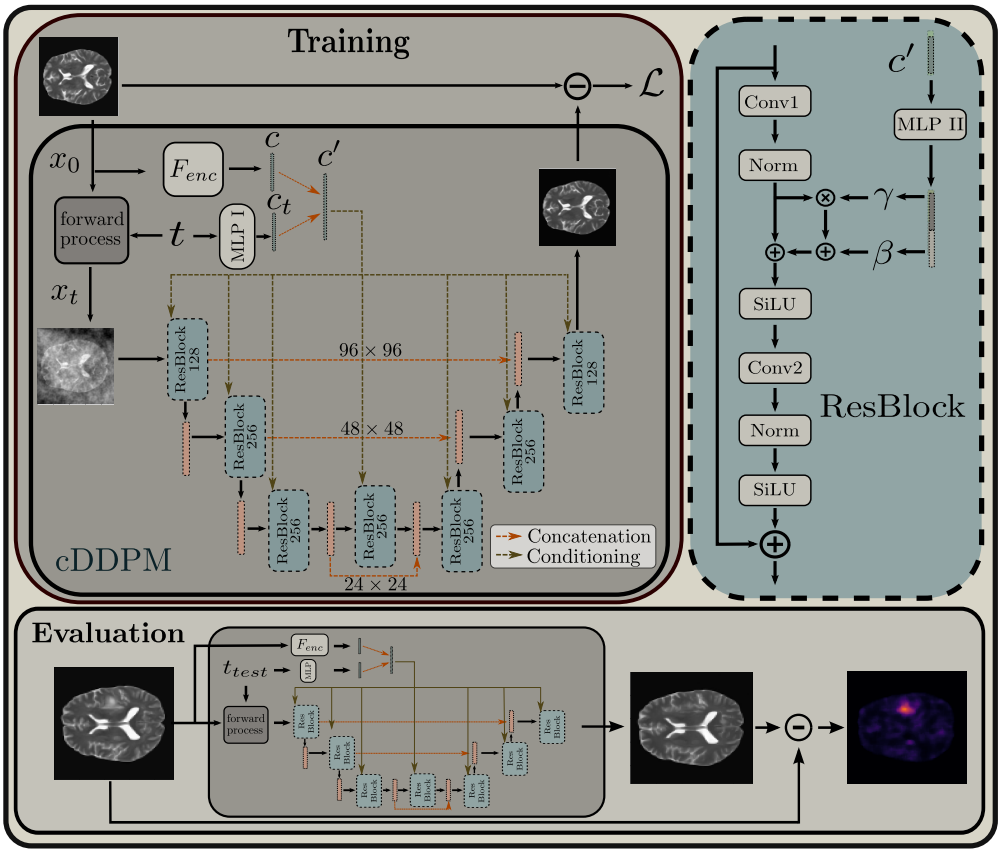


Fig. 3.4: Overview of our cDDPM, adapted from [22]. The timestep embedding c_t is concatenated with the projected encoder representation c of the input image, forming the conditioning vector c'. This vector is then used to scale and shift the feature maps within the residual blocks of the denoising Unet. Each residual block consists of two convolution layers (Conv1, Conv2), group normalization (Norm), and Sigmoid Linear Units (SiLU). During evaluation, the residual map is used for anomaly detection.

then utilize the context vector $\boldsymbol{c}$ to condition the denoising process. To incorporate $\boldsymbol{c}$ within the denoising Unet, we modify the timestep conditioning by concatenating $\boldsymbol{c}$ with the embedded timestep conditioning vector $\boldsymbol{c}_t \in \mathbb{R}^d$, resulting in a combined conditioning vector $\boldsymbol{c}' \in \mathbb{R}^{2 \cdot d}$. This vector $\boldsymbol{c}'$ is then used to adjust feature maps at each level of the Unet through a feature-wise scaling and shifting operation inspired by [147]. Specifically, an MLP generates two vectors, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, from $\boldsymbol{c}'$, with $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^{C_i}$, where C_i is the number of channels at Unet level i. The feature maps $\boldsymbol{f}_i$ at each level are transformed as:

$$\boldsymbol{f}_i' = \boldsymbol{f}_i \cdot (\boldsymbol{\gamma} + 1) + \boldsymbol{\beta}.$$

This transformation allows adaptive adjustment of feature maps based on the context vector $\boldsymbol{c}$ at each Unet level. Optionally, we pre-train the encoder network in a self-supervised manner using CNN-based masked image modeling [148].
At test time, the context vector $\boldsymbol{c}$ is derived from the input image and used to guide the denoising process. Similar to pDDPMs, we use the denoising Unet to generate a reconstruction from $\boldsymbol{x}_{t_{test}}$ with $t_{test} = \frac{T}{2} = 500$ in a single step. Additionally, we perform

an ablation study to assess the impact of different noise levels. Furthermore, we evaluate an ensemble approach using multiple noise levels $t_{\text{test}} = [250, 500, 750]$, where the final reconstruction is obtained by averaging the individual reconstructions from each noise level. The residual map between the input and the reconstructed image serves as the anomaly score.

3.4.4 Structural Similarity Index as Anomaly Score

This approach is based on our study *Diffusion Models with Ensembled Structure-based Anomaly Scoring for Unsupervised Anomaly Detection* [23], attached in Chapter 8.4.
Our previous studies demonstrated that both the type and architecture of the GM play a crucial role in UAD, as models with strong reconstruction capabilities can provide more accurate anomaly scoring. However, the selection of the discrepancy measure between input and reconstruction also directly impacts the accuracy of the segmentation. Commonly used metrics such as the l_1-error or the l_2-error focus on intensity-based discrepancies, which may fail to capture structural differences, as indicated by [80, 149]. This limitation may lead to missing subtle anomalies and overpenalizing errors from minor reconstruction imperfections.
We investigate using the SSIM as an alternative anomaly score. The SSIM incorporates structural, contrast, and luminance information based on local patch comparison rather than pixel-wise intensity comparison. The SSIM has demonstrated potential as an anomaly scoring method in reconstruction-based UAD [80, 142, 144]. However, while in [80], the evaluation is restricted to the domain of industrial defect detection, in the study [142], the SSIM is applied within the feature space of AEs and in [144] no comparison to other metrics, such as the $l1$-error, is presented. Furthermore, the application of SSIM in the context of DDPMs remains unexplored.
A pivotal element in the SSIM is the kernel dimension k_{ssim}, which defines the size of the compared local patches. Most existing approaches select a fixed kernel size based on the anticipated anomaly sizes, which may not be optimal for all pathologies. To address this issue, we propose an adaptive method (SSIM-ens) that computes a weighted average of SSIM scores across multiple kernel sizes. We hypothesize that this approach can reduce the dependency on a specific kernel size, enhancing robustness to different pathology types.

Approach

The SSIM between two images $\boldsymbol{x}$ and $\boldsymbol{y}$ is calculated by comparing local means, variances, and covariances [132]. These local statistics are computed by shifting a Gaussian kernel with spread σ across the images, where the kernel dimension is derived as $k_{\text{ssim}} = int(3.5 \cdot \sigma + 0.5) \cdot 2 + 1$, with int truncating the result to an integer.
SSIM scores are sensitive to the choice of σ, with larger values expanding the neighborhood of pixels considered and smaller values limiting it. To reduce dependency on a single scale, we propose SSIM-ens. This ensemble-based method averages SSIM scores across a range of σ values to provide a more robust anomaly score.
For a given input image $\boldsymbol{x}$ and its reconstruction $\boldsymbol{x}^{\text{rec}}$, the SSIM-ens anomaly score is computed as a weighted average over different σ values $\{\sigma_1, \sigma_2, \ldots, \sigma_n\}$:

$$\text{SSIM-ens}(\boldsymbol{x}, \boldsymbol{x}^{\text{rec}}) = 1 - \sum_{i=1}^{n} \boldsymbol{w}_i \cdot \text{SSIM}_{\sigma_i}(\boldsymbol{x}, \boldsymbol{x}^{\text{rec}}),$$

where

$$\boldsymbol{w}_i = \frac{e^{-\text{SSIM}_{\sigma_i}(\boldsymbol{x},\boldsymbol{x}^{\text{rec}})}}{\sum_{j=1}^{n} e^{-\text{SSIM}_{\sigma_j}(\boldsymbol{x},\boldsymbol{x}^{\text{rec}})}}.$$

Here, SSIM_{σ_i} denotes the SSIM score calculated with σ_i, and $\boldsymbol{w}_i$ is a normalized exponential weight inversely related to SSIM_{σ_i}, emphasizing regions with larger discrepancies. This ensemble method aims to improve robustness across varying pathology scales and types by combining multi-window SSIM scores. The SSIM-ens is then used as the final anomaly score.

3.4.5 Leveraging the Mahalanobis Distance to refine Anomaly Scoring

This approach is based on our study *Leveraging the Mahalanobis Distance to enhance Unsupervised Brain MRI Anomaly Detection* [24], attached in Chapter 8.5.
While SSIM can enhance anomaly scoring by capturing structural similarities, it is limited to assessing only the neighboring pixel context of individual inputs and reconstructions. Probabilistic models, however, enable the generation of multiple reconstructions, allowing us to model distributions that may provide additional contextual information. However, this information is seldom utilized in anomaly scoring, as the majority of existing methods either consider only a single reconstruction or average multiple reconstructions to derive the anomaly map [103, 62]. This approach ignores the contextual information that may be obtained from different reconstructions. Moreover, in studies that analyze multiple reconstructions, typically, only the variance across the reconstructions is considered, disregarding inter-pixel covariance [150, 151]. Therefore, we propose the Mahalanobis distance (MHD) [152] to capture the variability of pixel intensities within a pseudo-healthy distribution of multiple reconstructions to assess the deviation of individual pixels from a healthy reference distribution.
In our approach, we generate a pseudo-healthy distribution by reconstructing the same input image multiple times. The MHD is then computed pixel-wise between the input image and this distribution.
We hypothesize that using the MHD can enhance anomaly detection by differentiating genuine anomalies from common variations in pseudo-healthy reconstructions. Additionally, by capturing global spatial information, including potential symmetries across individual brain scans, the MHD could extend beyond the local neighborhood limitations of SSIM, leading to more accurate and robust anomaly segmentation.

Approach

Figure 3.5 illustrates the MHD anomaly scoring framework. To compute the MHD, we first generate a set of N pseudo-healthy reconstructions $\{\boldsymbol{x}^{\text{rec}_1}, \boldsymbol{x}^{\text{rec}_2}, \ldots, \boldsymbol{x}^{\text{rec}_N}\}$ for a given input image $\boldsymbol{x}$ using a cDDPM. We then calculate the mean reconstruction:

$$\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}^{\text{rec}_i}$$

and the covariance matrix:

$$\boldsymbol{\Sigma}_{\text{full}} = \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{x}^{\text{rec}_i} - \boldsymbol{\mu})(\boldsymbol{x}^{\text{rec}_i} - \boldsymbol{\mu})^\top \in \mathbb{R}^{H \cdot W \times H \cdot W},$$

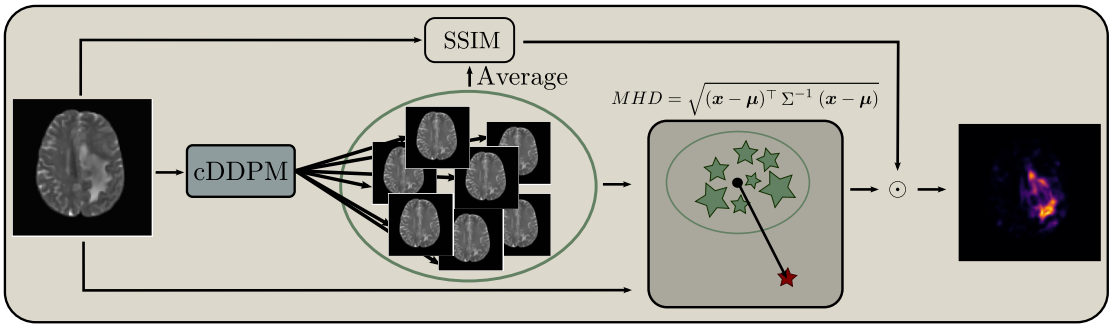


Fig. 3.5: Overview of or MHD-based anomaly scoring. First, a set of pseudo-healthy reconstructions is generated using a cDDPM. The mean reconstruction $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}_{\text{full}}$ are calculated across the reconstructions. The MHD is then computed pixel-wise between the input image $\boldsymbol{x}$ and the pseudo-healthy distribution. The final anomaly score is derived by element-wise multiplication of the SSIM and the MHD.

which captures the pixel-wise variance and covariance across the reconstructions.
The MHD between the input image $\boldsymbol{x}$ and the pseudo-healthy distribution is computed as:

$$\text{MHD}_{\text{full}}(\boldsymbol{x}) = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}_{\text{full}}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}.$$

This distance is calculated for each pixel, resulting in a pixel-wise anomaly score. Additionally, to analyze the effect of the covariances on the anomaly score, we also calculate the MHD using only the diagonal of the covariance matrix:

$$\text{MHD}_{\text{diag}}(\boldsymbol{x}) = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}_{\text{diag}}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}.$$

Here, $\boldsymbol{\Sigma}_{\text{diag}}$ is a diagonal matrix whose diagonal elements correspond to those of $\boldsymbol{\Sigma}_{\text{full}}$ with all off-diagonal elements set to zero.
To refine the anomaly map, we combine the MHD with the SSIM by calculating the SSIM between $\boldsymbol{x}$ and the mean reconstruction $\boldsymbol{\mu}$ and applying an element-wise product:

$$\text{S} = \text{MHD}(\boldsymbol{x}) \odot (1 - \text{SSIM}(\boldsymbol{x}, \boldsymbol{\mu})).$$

Notably, we compare different anomaly scores. S_{mean} denotes the averaging of multiple reconstructions to derive the anomaly map solely based on the SSIM. $\text{S}_{\text{diag}}^{MHD}$ and $\text{S}_{\text{full}}^{MHD}$ denote the use of the MHD either with a diagonal covariance matrix or with a full covariance matrix, respectively.

3.4.6 Supervised Anomaly Detection with Diffusion Models

This approach is based on our study *Combining Reconstruction-based Unsupervised Anomaly Detection with Supervised Segmentation for Brain MRIs* [25], attached in Chapter 8.6.
Reconstruction-based UAD methods can generalize to unseen pathologies but often produce noisy anomaly maps due to reconstruction artifacts. In contrast, supervised segmentation models can provide precise anomaly maps but require voxel-wise annotations and are limited to pathologies seen during training.

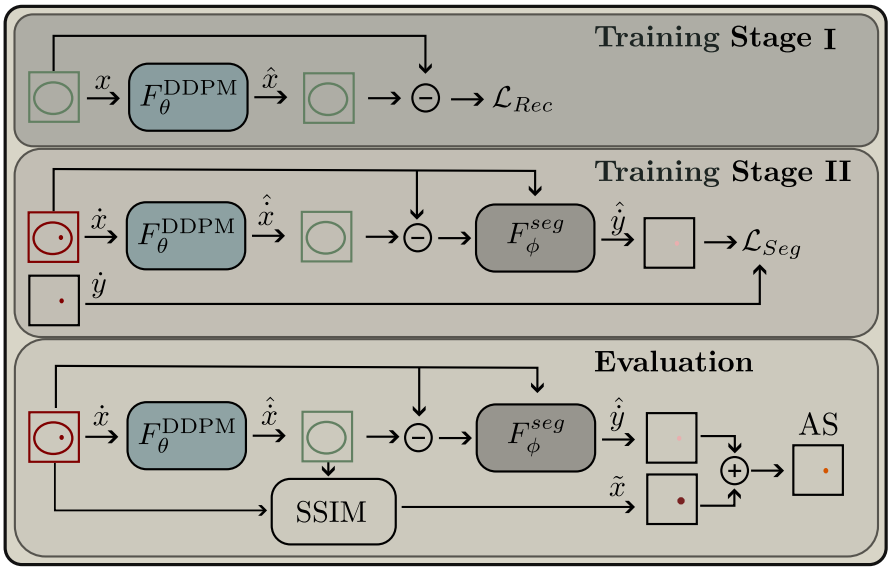


Fig. 3.6: Overview of our SADM framework, adapted from [25]. The framework consists of two branches: a DDPM for pseudo-healthy reconstructions and a supervised Unet for segmentation. In stage I, The DDPM is trained to reconstruct healthy brain scans. In stage II, discrepancies between the input and reconstruction are provided to the Unet to predict the segmentation map. The final anomaly score combines the unsupervised SSIM with the supervised segmentation prediction.

To combine the strengths of both approaches, we propose a two-branch framework that integrates the general anomaly detection capabilities of reconstruction-based methods with the precision of supervised segmentation. First, we train a GM to reconstruct healthy images, producing pseudo-healthy reconstructions for input images with potential anomalies. Subsequently, we train a supervised segmentation model with the goal of identifying anomalies focusing on deviations between the input and its reconstruction. At test time, we combine the anomaly maps produced by the reconstruction branch with the segmentation predictions from the supervised branch, resulting in a combined anomaly score.
We hypothesize that this framework improves segmentation accuracy for known pathologies while maintaining generalizability to novel anomalies. Furthermore, additional information is fed to the supervised segmentation model by focusing on the residual of input and pseudo-healthy reconstruction. This additional information potentially enhances the generalization beyond the anomalies used in the training data.

Approach

An illustration of the proposed approach is illustrated in Figure 3.6. Our Supervised Anomaly Detection with Diffusion Models (SADM) framework integrates two primary branches: a DDPM for pseudo-healthy reconstructions (reconstruction branch) and a supervised Unet for segmentation (segmentation branch). The SADM training process consists of two sequential stages:
Stage I: Unsupervised Reconstruction: In the first stage, the DDPM is trained to reconstruct healthy brain scans following a reconstruction-based UAD strategy. The

DDPM is optimized to minimize the reconstruction error $\mathcal{L}_{Rec} = |\boldsymbol{x} - \boldsymbol{x}^{\text{rec}}|$, where $\boldsymbol{x}^{\text{rec}} = F_\theta^{\text{DDPM}}(\boldsymbol{x})$ denotes the pseudo-healthy reconstruction of the input scan $\boldsymbol{x}$. In our experiments, we use our cDDPM for F_θ^{DDPM}.

Stage II: Supervised Segmentation: In the second stage, the pseudo-healthy reconstructions from stage I are used to support anomaly segmentation. For an abnormal input scan $\dot{\boldsymbol{x}}$ with ground truth annotation $\dot{\boldsymbol{y}}$, we generate a pseudo-healthy reconstruction $\dot{\boldsymbol{x}}^{rec} = F_\theta^{\text{DDPM}}(\dot{\boldsymbol{x}})$. The difference $(\dot{\boldsymbol{x}} - \dot{\boldsymbol{x}}^{rec})$ and the original input $\dot{\boldsymbol{x}}$ are provided to a Unet to predict the segmentation map:

$$\hat{\dot{\boldsymbol{y}}} = F_\phi^{\text{seg}}(\dot{\boldsymbol{x}} - \dot{\boldsymbol{x}}^{rec}, \dot{\boldsymbol{x}}).$$

Both $\dot{\boldsymbol{x}} - \dot{\boldsymbol{x}}^{rec}$ and $\dot{\boldsymbol{x}}$, are encoded by the encoder of the Unet. Subsequently, the resulting feature maps are concatenated at each layer and fed to the Unet decoder. The Unet is trained by minimizing the cross-entropy loss for segmentation $\mathcal{L}_{\text{Seg}} = CE(\hat{\dot{\boldsymbol{y}}}, \dot{\boldsymbol{y}})$. During this stage, the DDPM parameters θ are frozen. For anomaly detection, both branches of the SADM framework are utilized. Given a potentially abnormal input $\dot{\boldsymbol{x}}$, the DDPM generates a pseudo-healthy reconstruction $\dot{\boldsymbol{x}}^{rec} = F_\theta^{\text{DDPM}}(\dot{\boldsymbol{x}})$. The supervised segmentation network then predicts the anomaly map $\hat{\dot{\boldsymbol{y}}} = F_\phi^{\text{seg}}(\dot{\boldsymbol{x}} - \dot{\boldsymbol{x}}^{rec}, \dot{\boldsymbol{x}})$. Additionally, we compute the pixel-wise SSIM between the input and reconstruction for unsupervised anomaly scoring:

$$\tilde{\boldsymbol{x}} = 1 - SSIM(\dot{\boldsymbol{x}}, \dot{\boldsymbol{x}}^{rec}).$$

The final Anomaly Score (AS) combines the unsupervised anomaly map with the supervised segmentation prediction:

$$\mathbf{AS} = \tilde{\boldsymbol{x}} + \hat{\dot{\boldsymbol{y}}}.$$

This combined approach aims to leverage the broad but general anomaly detection capability of the unsupervised anomaly map along with the precise segmentation of known anomalies provided by the supervised network.

4 Experimental Results

Tab. 4.1: Results for 2D and 3D VAEs combined with spatial erasing strategies, adapted from [20]. DICE ($\mu \pm \sigma$) represents the mean and standard deviation of the DICE scores across subjects. The AUPRC is computed by concatenating predictions and annotations across the entire test set. The highest values are highlighted in bold, and all metrics are reported as percentages.

	BRATS (T2)		ATLAS (T1)	
Input & Erasing	**DICE**	**AUPRC**	**DICE**	**AUPRC**
2D-None	25.30 ± 12.37	21.19	11.23 ± 13.66	16.86
3D-None	26.93 ± 12.40	24.69	14.42 ± 16.06	23.74
2D-Patch-0	26.52 ± 13.42	22.53	12.23 ± 13.67	18.65
2D-Patch-n	26.58 ± 13.27	22.54	12.36 ± 14.61	18.20
3D-Cube-0	27.90 ± 13.57	26.18	**15.59 ± 17.02**	23.47
3D-Cube-n	**28.80 ± 13.74**	**27.85**	15.53 ± 17.30	25.11
2D-Multi-Patch-0	26.44 ± 12.89	22.54	11.82 ± 14.29	18.72
2D-Multi-Patch-n	27.24 ± 13.14	22.81	12.88 ± 15.21	19.49
3D-Multi-Cube-0	27.67 ± 13.22	25.82	15.23 ± 16.64	24.51
3D-Multi-Cube-n	28.33 ± 13.42	26.18	14.99 ± 17.31	25.13
2D-Half-Slice-0	25.44 ± 12.42	21.77	11.05 ± 13.70	18.60
2D-Half-Slice-n	26.45 ± 13.22	22.84	12.13 ± 14.79	20.37
3D-Half-Volume-0	27.51 ± 13.17	25.47	15.21 ± 17.00	23.14
3D-Half-Volume-n	27.92 ± 13.24	26.07	15.27 ± 17.21	**25.58**

In this chapter, we summarize the results of our proposed methods and experiments. For additional results, ablation studies and visualizations, we refer to the corresponding publications, attached in Chapter 8.

4.1 3D VAEs for UAD in Brain MRI

We present the results of our paper, *3-Dimensional Deep Learning with Spatial Erasing for Unsupervised Anomaly Segmentation in Brain MRI* [20], attached in Chapter 8.1. All models are trained on T1-weighted scans from the MN data set and evaluated on the BRATS 2019 and ATLAS v1 data sets. For evaluation, we report the mean and standard deviation of the DICE score across subjects, with thresholds optimized on a held-out validation set. Additionally, we calculate the AUPRC by concatenating predictions and annotations across the entire test set to provide a threshold-independent metric.
We compare 2D and 3D VAEs and evaluate different erasing strategies. These strategies

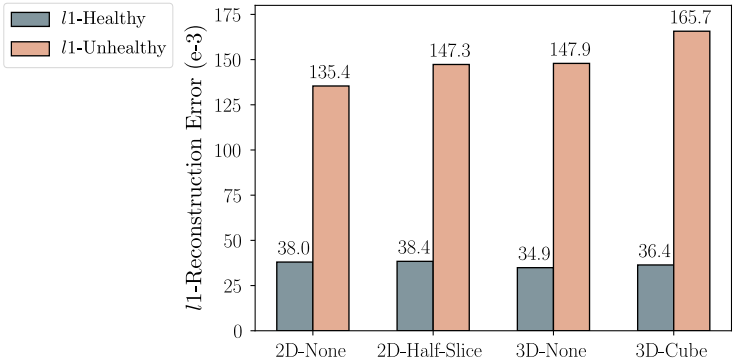


Fig. 4.1: $l1$-error for reconstructions of healthy and unhealthy brain regions for 2D and 3D VAEs with different erasing strategies. Healthy and unhealthy regions are defined by the ground truth annotations from the BRATS data set.

include *patch*, *multi-patch*, and *half-slice* for 2D models, and *cube*, *multi-cube*, and *half-volume* for 3D models. Each erasing strategy is tested by replacing the erased regions with zeros (*-0) and noise (*-n). The results are summarized in Table 4.1.
Our findings show that 3D VAEs consistently outperform their 2D counterparts, in terms of DICE and AUPRC. Furthermore, we observe improved performance when applying erasing. While there is no clear superiority of one erasing strategy over the others, the *Cube* and *Half-Volume* methods for 3D models demonstrate robust performance. Across the different inpainting strategies (noise or zeros), similar performance is reported with a slight benefit for noise.
In Figure 4.1, we present a comparison of the $l1$-error between healthy and unhealthy regions. Healthy regions exhibit lower $l1$-error values compared to unhealthy brain structures. Particularly for erasing strategies, the errors are substantially higher in unhealthy regions while only marginally increased in healthy regions.
Overall, 3D processing can improve the segmentation performance and reconstruction accuracy. However, with VAEs, the reconstructions remain blurry and generic, and the segmentation performance is moderate, as also illustrated in Figure 4.3.

4.2 Diffusion Models for UAD in Brain MRI

We summarize the results of our papers *Patched Diffusion Models for Unsupervised Anomaly Detection in Brain MRI* [21] and *Guided Reconstruction with Conditioned Diffusion Models for Unsupervised Anomaly Detection in Brain MRIs* [22], attached in Chapter 8.2 and 8.3. All models are trained on T1- and T2-weighted scans from the IXI data set and evaluated on BRATS 2021 (T2), MSLUB (T2), ATLAS v2 (T1), and WMH (T1) data sets. The metrics are averaged across five cross-validation folds. The segmentation performance is evaluated using the Dice score ($\lceil$DICE$\rceil$). The reconstruction quality is assessed using SSIM, LPIPS, PSNR, and $l1$-error on healthy data. Furthermore, the ratio of the $l1$-error on unhealthy and healthy data is considered ($l1$-ratio).

Tab. 4.2: Comparison of the reconstruction quality of the different models for the healthy IXI data set, adapted from [22]. The highest values are shown in **bold**, where underlines denote statistical significance ($p < 0.05$). For all metrics, the mean ± standard deviation across the different folds are reported. The arrows ↑ and ↓ indicate that higher and lower values are favorable, respectively.

Model	**SSIM** ↑	**PSNR** ↑	**LPIPS** (e-3) ↓	***l*1-error** (e-3) ↓
VAE [62]	74.98 ± 0.54	23.38 ± 0.14	4.03 ± 0.50	32.32 ± 0.64
SVAE [137]	77.87 ± 0.15	23.94 ± 0.06	3.31 ± 0.24	29.08 ± 0.16
AE [62]	76.11 ± 0.27	23.41 ± 0.14	3.19 ± 0.54	31.67 ± 0.41
RA [138]	75.46 ± 0.35	23.92 ± 0.23	2.18 ± 0.41	34.36 ± 1.43
PHANES [140]	69.04 ± 1.23	21.39 ± 0.32	1.08 ± 0.09	38.7 ± 1.74
DAE [115]	**98.69 ± 0.15**	**36.69 ± 0.38**	0.14 ± 0.01	**8.14 ± 0.17**
AnoDDPM [116]	93.96 ± 0.37	31.79 ± 0.26	0.49 ± 0.14	14.29 ± 0.32
pDDPM [21]	96.62 ± 0.25	34.58 ± 0.39	**0.09 ± 0.04**	9.70 ± 0.43
cDDPM [22]	96.80 ± 0.19	34.87 ± 0.23	0.11 ± 0.05	9.68 ± 0.16

Reconstruction Quality In the previous section, we demonstrated that 3D VAEs outperform their 2D counterparts in reconstruction quality, achieving an 8.8 % reduction in reconstruction error for healthy brain regions, as illustrated in Figure 4.1. Considering DDPM-based approaches, already the baseline DDPM (AnoDDPM) significantly outperforms 2D VAEs, achieving a remarkable 126.2% reduction in $l1$-error ($p < 0.05$), according to Table 4.2. Moreover, our proposed extensions, namely pDDPM and cDDPM, further enhance performance, outperforming AnoDDPM in reconstruction quality with statistically significant improvements ($p < 0.05$). Our results highlight that models using Unet-like architectures with denoising tasks for regularization, such as DDPMs or DAEs, consistently demonstrate significantly higher reconstruction quality than dense AEs or VAEs. To evaluate the reconstruction capabilities considering pathological structures, we examine the unhealthy-to-healthy error ratio ($l1$-ratio) on the unhealthy test sets, as presented in Table 4.3. A higher $l1$-ratio is preferable, as it indicates the model's capacity to reconstruct healthy anatomy without replicating unhealthy regions. A lower $l1$-ratio indicates that the model tends to replicate unhealthy structures in the reconstruction or fails to reconstruct healthy anatomy. Although DAEs achieve the lowest $l1$-error among all models, their $l1$-ratio is moderate, particularly for the MSLUB and ATLAS data sets. These results demonstrate that DAEs tend to replicate unhealthy structures. In contrast, cDDPMs achieve the highest $l1$-ratio across most data sets, with the exception of the WMH data set, where AnoDDPM demonstrates competitive performance.

Segmentation Performance As illustrated in Table 4.4, DDPM-based approaches yield significant improvements in segmentation performance. The AnoDDPM baseline exhibits a Dice score enhancement of up to 51.8% compared to VAEs. The proposed pDDPM demonstrates further performance improvements, considering T2-weighted data. Moreover, our cDDPM model exhibits superior segmentation performance across most data sets. The performance of cDDPMs is further enhanced by pre-training the encoder (SSL checkmark) and by ensembling reconstructions from different values of t_{test} (ENS checkmark).

Tab. 4.3: Comparison of the $l1$-ratio for healthy and unhealthy brain regions, adapted from [22]. The highest values are shown in **bold**. For all metrics, the mean ± standard deviation across the different folds are reported. The arrows ↑ and ↓ indicate that higher and lower values are favorable, respectively.

Model	BRATS (T2) ***l*1-ratio** ↑	MSLUB (T2) ***l*1-ratio** ↑	ATLAS (T1) ***l*1-ratio** ↑	WMH (T1) ***l*1-ratio** ↑
VAE [62]	3.52 ± 0.08	2.92 ± 0.06	4.43 ± 0.03	2.36 ± 0.04
SVAE [137]	3.90 ± 0.05	3.13 ± 0.05	3.38 ± 0.11	2.07 ± 0.01
AE [62]	3.84 ± 0.17	3.26 ± 0.18	4.40 ± 0.07	2.36 ± 0.04
RA [138]	3.10 ± 0.16	2.56 ± 0.11	3.93 ± 0.25	2.42 ± 0.19
PHANES [140]	3.54 ± 0.13	2.73 ± 0.07	4.01 ± 0.07	2.20 ± 0.05
DAE [115]	7.17 ± 0.63	2.69 ± 0.15	4.51 ± 0.15	2.99 ± 0.14
AnoDDPM [116]	6.16 ± 0.53	3.37 ± 0.24	5.00 ± 0.23	**3.16 ± 0.15**
pDDPM [21]	7.16 ± 0.15	4.34 ± 0.13	5.58 ± 0.28	3.00 ± 0.16
cDDPM [22]	**7.43 ± 0.17**	**4.49 ± 0.18**	**5.69 ± 0.27**	3.12 ± 0.08

As illustrated in Figure 4.2, the performance of all DDPM-based models is sensitive to the degree of noise introduced during the forward process. The pDDPM and cDDPM models demonstrate improvements in segmentation performance over AnoDDPM across varying noise levels t_{test}, except for the WMH data set, where improvements are observed only at high noise levels $t_{test} > 500$. Additionally, ensembling different noise levels for the cDDPM leads to consistently high performance across all data sets, thereby mitigating the impact of selecting individual noise levels.

Besides improved performance metrics, compared to pDDPMs, cDDPMs eliminate the need to select a specific patch size, thereby enhancing generalization. Moreover, cDDPMs exhibit a reduction in inference time of approximately 37% compared to pDDPMs, with only a slight increase of roughly 2% compared to AnoDDPM. These results were obtained on an NVIDIA RTX 3090 GPU.

Figure 4.3 presents a qualitative analysis of exemplary reconstructions and anomaly maps for VAE, AnoDDPM, pDDPM and cDDPM models. VAEs fail to accurately reconstruct the input image, leading to poor segmentation performance. The AnoDDPM shows improved reconstruction quality. However, the intensity distribution of the input image is not captured accurately, leading to false positive predictions across the entire anomaly map. In contrast, pDDPM and cDDPM, in addition to accurate reconstructions, better capture the input image's intensity distribution, as shown in the provided histograms. The anomaly maps generated by the cDDPM provide the strongest contrast between healthy and unhealthy regions, reducing false positive predictions in the final thresholded segmentation.

Among the non-DDPM baselines presented in Table 4.4, models such as FAE, PII, and DAE exhibit comparable performance compared to DDPM-based methods, considering the BRATS data set. However, their performance is inconsistent across other data sets. Similarly, the RD, PHANES, and EDC models demonstrate competitive performance on the ATLAS data set but fail to generalize across other data sets. The threshold-based method, Thresh, demonstrates limited performance across all data sets. These results

Tab. 4.4: Comparison of the segmentation performance regarding ⌈DICE⌉, adapted from [22]. The highest values are shown in **bold**, where underlines denote statistical significance ($p < 0.05$). For all metrics, the mean ± standard deviation across the different folds are reported. A checkmark at SSL denotes that a pre-trained encoder is used. A checkmark at ENS denotes the ensembling of different values for $t_{test} \in [250, 500, 750]$. Otherwise, a fixed value of $t_{test} = 500$ is used for DDPM-based models.

Model	Modification ENS	SSL	BRATS (T2) ⌈**DICE**⌉	ATLAS (T1) ⌈**DICE**⌉	MSLUB (T2) ⌈**DICE**⌉	WMH (T1) ⌈**DICE**⌉
Thresh [142]			30.26	4.66	7.65	10.32
VAE [62]			33.12 ± 1.12	15.63 ± 0.73	8.10 ± 0.18	7.60 ± 0.28
SVAE [137]			36.43 ± 0.36	10.32 ± 0.53	8.55 ± 0.11	7.18 ± 0.07
AE [62]			36.04 ± 1.73	14.04 ± 0.60	9.65 ± 0.97	7.34 ± 0.08
DAE [115]			48.82 ± 3.68	15.95 ± 0.69	7.57 ± 0.61	12.02 ± 1.01
RA [138]			16.75 ± 0.51	12.21 ± 0.98	3.96 ± 0.03	6.04 ± 0.45
PHANES [140]			28.42 ± 0.91	17.62 ± 0.41	6.11 ± 0.27	7.55 ± 0.17
RD [82]			32.57 ± 0.15	19.69 ± 0.26	6.48 ± 0.20	7.48 ± 0.10
FAE [101]			44.59 ± 2.19	17.76 ± 0.16	6.85 ± 0.65	8.81 ± 0.38
EDC [100]			36.66 ± 3.03	18.67 ± 1.02	7.23 ± 0.29	8.62 ± 0.47
PII [119]			40.83 ± 2.18	9.73 ± 1.89	9.46 ± 0.43	6.59 ± 1.87
AnoDDPM [116]			49.43 ± 1.94	17.57 ± 1.05	9.63 ± 1.33	11.56 ± 0.93
AnoDDPM [116]	✓		50.27 ± 2.67	20.18 ± 0.58	9.71 ± 1.29	**12.06 ± 0.97**
pDDPM [21]			53.25 ± 0.50	19.20 ± 0.45	12.40 ± 0.36	10.14 ± 0.50
pDDPM [21]	✓		53.61 ± 0.51	19.92 ± 0.24	12.83 ± 0.40	10.13 ± 0.53
cDDPM [22]			54.49 ± 1.63	22.6 ± 1.67	12.79 ± 1.08	11.21 ± 0.54
cDDPM [22]		✓	55.67 ± 1.05	22.66 ± 1.20	13.52 ± 0.91	11.15 ± 0.8
cDDPM [22]	✓	✓	**56.30 ± 1.25**	**24.22 ± 1.10**	**14.04 ± 1.16**	11.59 ± 0.93

demonstrate the challenging task of generalization for UAD methods and highlight the generalization capabilities of our cDDPM model, which consistently outperforms baseline methods and shows robustness across diverse data sets.

4.3 Enhancing Anomaly Scoring for UAD in Brain MRI

This section presents a summary of our papers *Diffusion Models with Ensembled Structure-based Anomaly Scoring for Unsupervised Anomaly Detection* [23] and *Leveraging the Mahalanobis Distance to enhance Unsupervised Brain MRI Anomaly Detection* [24], attached in Chapter 8.4 and 8.5.

Initially, we investigate the use of SSIM in conjunction with cDDPMs, with a particular focus on how the σ parameter of SSIM influences segmentation performance. The σ parameter represents the standard deviation of the Gaussian kernel, which determines the window size for computing local statistics of the SSIM. While in our original paper, we focus on AnoDDPM, in Table 4.5 and Figure 4.4, we extend the investigation to cDDPMs to enable a direct comparison. The models are trained on the BRATS 2021 (T2), MSLUB (T2), ATLAS v2 (T1), and WMH (T1) data sets. The segmentation performance is evaluated using the Dice score (⌈DICE⌉), averaged across five cross-validation folds.

As illustrated in Figure 4.4, the σ parameter influences segmentation performance, showing diverging trends across the compared data sets. As a result, determining the

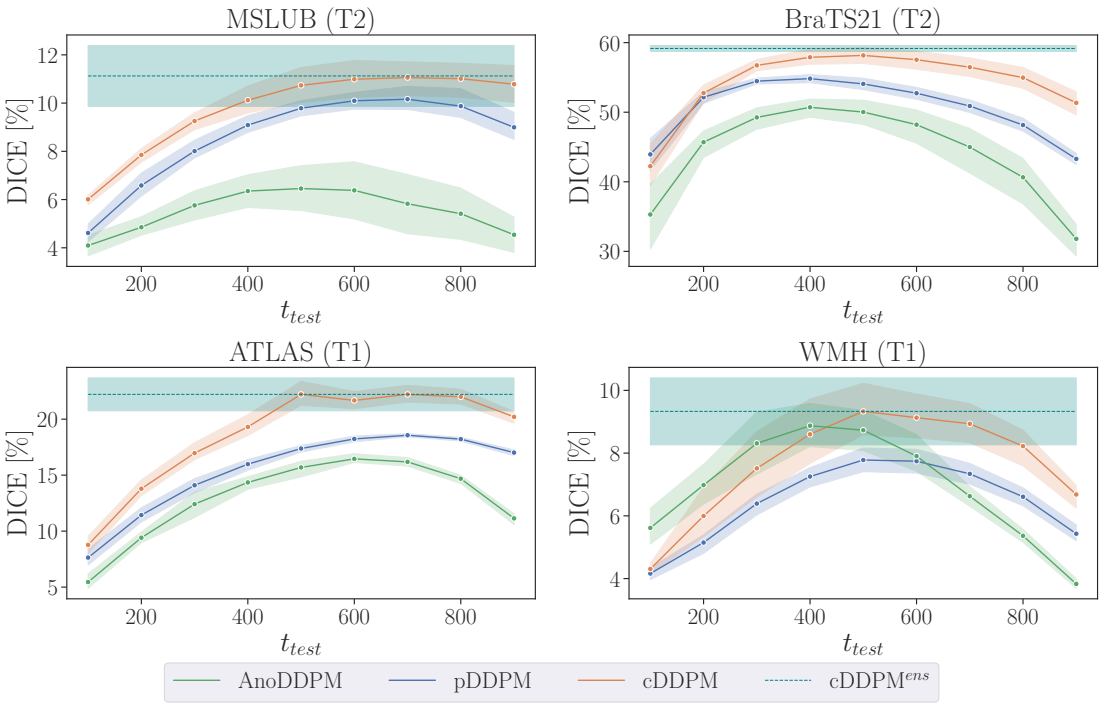


Fig. 4.2: Segmentation performance for different noise levels t_{test}, adapted from [22]. Top row: MSLUB (left) and BRATS (right) data sets. Bottom row: ATLAS (left) and WMH (right) data sets. The superscript *ens* denotes the ensembling of reconstructions from different noise levels $t_{test} \in [250, 500, 750]$.

optimal value of σ is challenging, as it varies across different data sets. For instance, in the MSLUB data set, a value of $\sigma = 0.3$ yields optimal performance, whereas for the BRATS data set, $\sigma = 1.6$ is more suitable. SSIM-ens provides robust and consistent performance across different data sets by taking weighted averages of the anomaly scores across a range of σ values. While it does not always surpass the performance of the optimal σ_{best} value, its adaptive ensemble approach mitigates the need for manual σ selection, offering a more generalized and stable performance across data sets comprising different pathologies.
Table 4.5 demonstrates that using SSIM-ens as an anomaly score for cDDPMs improves segmentation performance compared to the $l1$-error and the SSIM with $\sigma = 1$. Moreover, the results emphasize the influence of anomaly scoring strategies on segmentation performance. Transitioning from AnoDDPM to cDDPM leads to an average performance improvement of 24%, while replacing the $l1$-error with SSIM-ens further enhances performance by an average of 16%. These findings highlight the importance of selecting appropriate anomaly scoring strategies for UAD in brain MRI, particularly in reconstruction-based approaches.

Building on these results, we investigate the use of the MHD to improve anomaly scoring further. In contrast to the original paper, where no post-processing was applied, we present the results with post-processing to enable a direct comparison. We consider cDDPMs to generate the pseudo-healthy distributions for the MHD calculation. To achieve an optimal balance between performance and inference time, we generate $N = 10$

reconstructions for each input image and employ the SSIM with $\sigma = 1.0$ as the anomaly score.
According to the results in Table 4.5, averaging multiple reconstructions of cDDPMs (S_{mean}) already enhances segmentation performance across the majority of data sets compared to the use of a single reconstruction. Furthermore, the application of the MHD with a full covariance matrix (S_{full}^{MHD}) consistently exhibits enhanced performance across all data sets. In contrast, the application of the MHD with a diagonal covariance matrix (S_{diag}^{MHD}) does not improve the segmentation performance compared to averaging (S_{mean}). This observation indicates that the full covariance matrix is important to enhance the segmentation performance. Figure 4.5 (a) compares the anomaly maps derived from S_{mean}, the isolated MHD_{diag} and MHD_{full}. Compared to S_{mean}, both MHD variants can effectively reduce false positive predictions. Additionally, incorporating the full covariance matrix preserves anomaly edges better than the diagonal covariance approach. The multiplication of S_{mean} with MHD_{full} results in the most accurate anomaly map, reducing false positives and enhancing the segmentation performance. Figure 4.5 (b) illustrates the correlation of a single pixel (indicated by the green arrow) with all other pixels in the image. Such dependencies can only be captured by the full covariance

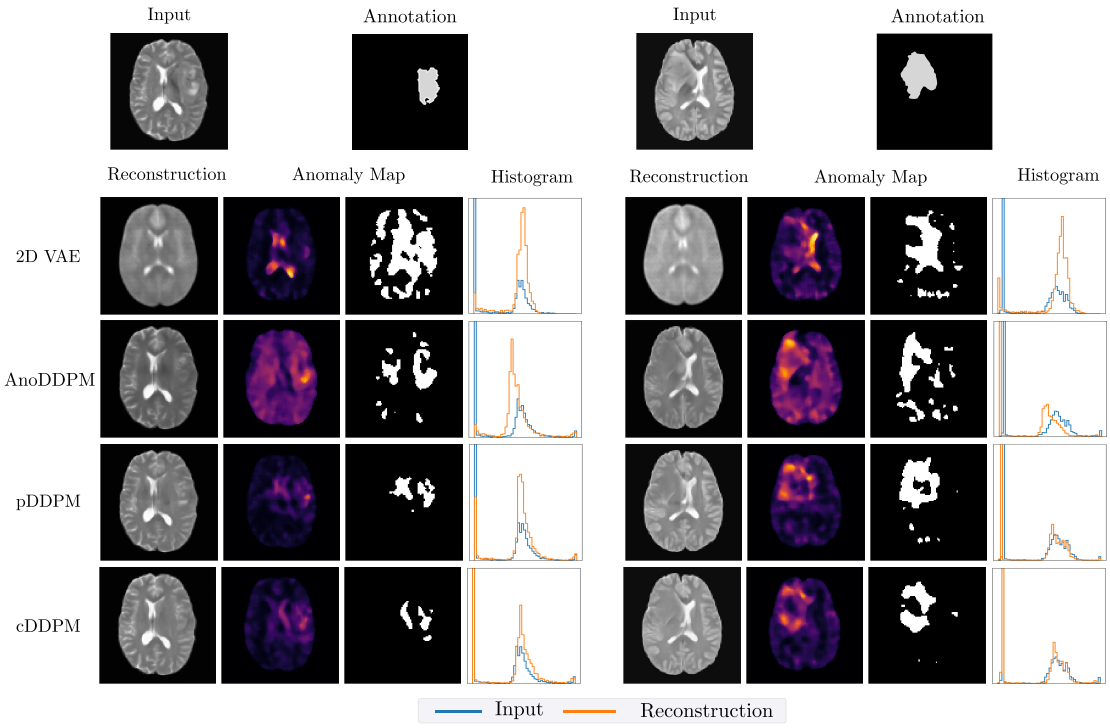


Fig. 4.3: Exemplary reconstructions and anomaly maps for 2D VAE, AnoDDPM, pDDPM and cDDPM models, adapted from [22]. The input and the corresponding ground truth annotation are provided in the first row. For each case, the reconstruction, the anomaly map and the histograms of intensity values in input and reconstruction are shown. Note that for histogram calculation, only healthy areas are considered. For visualization purposes, we provide segmentation maps next to the anomaly maps. We derive the binarization threshold by optimizing it for the best possible Dice score.

Tab. 4.5: Segmentation performance regarding ⌈DICE⌉. The highest values are shown in **bold**, where underlines denote statistical significance ($p < 0.05$). S_{mean} denotes the averaging of multiple reconstructions to derive the anomaly map. S^{MHD}_{diag} and S^{MHD}_{full} denote the use of the MHD either with a diagonal covariance matrix or with a full covariance matrix, respectively.

Model	BRATS (T2) ⌈**DICE**⌉	ATLAS (T1) ⌈**DICE**⌉	MSLUB (T2) ⌈**DICE**⌉	WMH (T1) ⌈**DICE**⌉
cDDPM ($l1$)	56.30 ± 1.25	24.22 ± 1.10	14.04 ± 1.16	11.59 ± 0.93
cDDPM (SSIM)	65.78 ± 0.58	23.88 ± 1.23	13.34 ± 0.46	16.91 ± 0.87
cDDPM (SSIM-ens)	67.30 ± 0.49	25.07 ± 1.29	13.93 ± 0.43	16.55 ± 0.88
cDDPM S_{mean}	68.19 ± 0.35	24.16 ± 1.37	13.29 ± 0.61	17.18 ± 1.25
cDDPM S^{MHD}_{diag}	68.57 ± 0.35	24.15 ± 1.44	13.64 ± 0.62	17.53 ± 1.51
cDDPM S^{MHD}_{full}	**70.55 ± 0.28**	**27.73 ± 1.67**	**15.55 ± 0.83**	**17.79 ± 1.70**

matrix and are disregarded by the diagonal covariance matrix, resulting in the loss of information.
In summary, the anomaly scoring method substantially impacts the performance of the evaluated GMs. Our results demonstrate that using the SSIM as an anomaly score can enhance the segmentation performance of cDDPMs. Moreover, complementing the SSIM with the MHD can significantly improve the segmentation performance of cDDPMs.

4.4 Supervised Anomaly Detection with Diffusion Models

We present the results of our paper *Combining Reconstruction-based Unsupervised Anomaly Detection with Supervised Segmentation for Brain MRIs* [25], attached in Chapter 8.6. Compared to previous approaches, the training process is divided into distinct stages for SADM, as illustrated in Figure 3.6. All training stages and evaluations are conducted using T1-weighted MRI scans. In stage I, the GMs are trained using the IXI data set. In Stage II, a strategy to generate pairs of synthetic anomalies and ground truth annotation presented in [141] is employed based on the IXI data set. The resulting

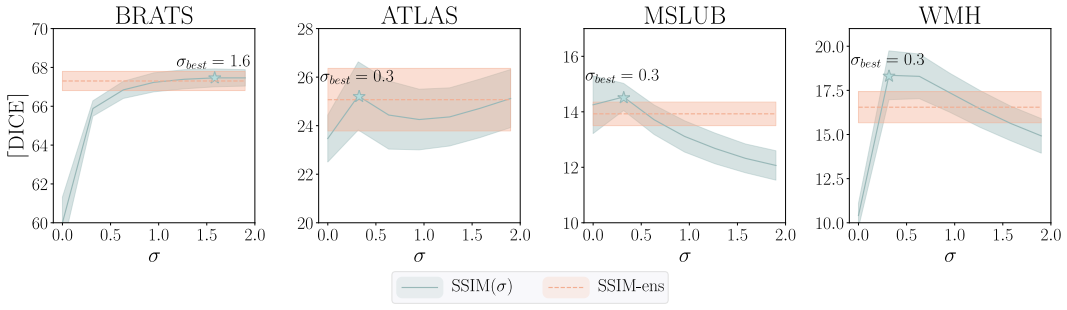


Fig. 4.4: Comparison of the segmentation performance for varying σ values. Optimal performance is denoted by (σ_{best}). The center lines delineate the mean performance, while the surrounding shaded regions depict the standard deviation across five folds. Dashed lines indicate the performance of SSIM-ens.

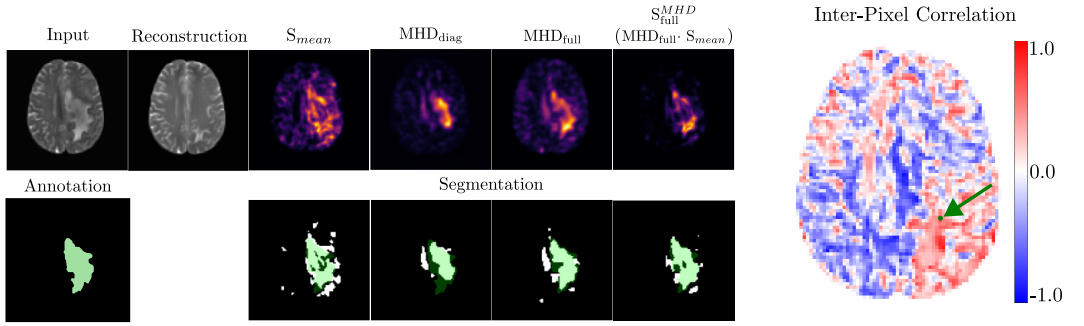


Fig. 4.5: Qualitative results of the Mahalanobis distance, adapted from [24]. (a): **Top row:** input, reconstruction, S_{mean}, MHD_{diag}, MHD_{full} and (S^{MHD}_{full}) are shown for an exemplary image taken from the BRATS data set. **Bottom row:** The ground truth (green) and binarized segmentation maps (white) are shown. Note that the threshold for the segmentation maps is derived by optimizing the Dice score based on the ground truth. (b): The correlation of one pixel (green arrow) with all other pixels, derived from Σ_{full}, is visualized as a heatmap.

data set is referred to as DRAEM. Furthermore, small subsets comprising approximately 10% of the BRATS 2021 and ATLAS v2 data sets are utilized in the context of semi supervised learning. For evaluation, we utilize the remaining samples of the BRATS and ATLAS data sets, respectively. Furthermore, we utilize the augmented IXI test set (DRAEM) to assess the segmentation performance concerning synthetic anomalies. The results are presented in Table 4.6.

Different variants of SADM are evaluated. In $SADM_{res}$, the residual of the input and the reconstruction, along with the input, is fed to the Unet. In contrast, only the input is used in SADM. Moreover, we examine the Unet and $Unet_{res}$ variants. In contrast to SADM, only the Unet prediction is utilized in these models, while the anomaly map produced by the unsupervised reconstruction branch is disregarded.

Initially, we consider the typical UAD scenario, where only data with healthy labels is available. In this case, synthetic anomalies are employed to generate a supervised signal for the segmentation branch in SADM. Our framework is then compared to several UAD baselines in this setting. The results are presented in blocks I and II of Table 4.6. Across the compared UAD baselines in block I, cDDPMs show the highest segmentation performance for real pathologies. Hence, we consider them as a reconstruction model for the SADM framework. Next, we consider self supervised training with the synthetic anomalies in DRAEM (block II). In this setting, $SADM_{res}$ outperforms cDDPMs with performance improvements of 3.4 % and 12.3 % for the BRATS and ATLAS data sets, respectively. Furthermore, $SADM_{res}$ demonstrates a substantial improvement of 542.5 % in segmentation performance for synthetic anomalies in the DRAEM data set. Next, we investigate using our framework in a semi supervised setting. Instead of generating synthetic anomalies, we assume that a small amount of annotated data is available and consider subsets of the BRATS and ATLAS data sets for training. We limit training to a single data set at a time to assess the model's ability to generalize to unseen pathologies. The results for this semi supervised setting are presented in block III of Table 4.6. Using a limited set of annotated data leads to notable improvements in the segmentation

Tab. 4.6: Segmentation performance for supervised anomaly detection with diffusion models, adapted from [25]. **Block I:** Unsupervised approaches, trained with healthy data. **Block II:** Self-supervised approaches, trained with synthetic anomalies. **Block III:** Semi supervised approaches, trained with real pathologies. $\mathcal{D}_{healthy}$ and $\mathcal{D}_{unhealthy}$ represent the type of data used during training.

		Training Data		**Test Data**		
	Model	$\mathcal{D}_{healthy}$	$\mathcal{D}_{unhealthy}$	BRATS (real)	ATLAS (real)	DRAEM (synthetic)
I. Unsupervised	AE [62]	IXI	None	39.16 ± 0.64	14.14 ± 0.28	9.91 ± 0.04
	VAE [62]	IXI	None	39.25 ± 0.50	14.52 ± 0.37	9.83 ± 0.14
	DAE [115]	IXI	None	55.93 ± 0.66	19.95 ± 0.96	12.50 ± 0.31
	FAE [101]	IXI	None	43.04 ± 0.49	17.59 ± 0.15	**19.60 ± 0.49**
	RD [82]	IXI	None	32.90 ± 0.65	19.45 ± 0.25	19.55 ± 0.60
	AnoDDPM [116]	IXI	None	48.65 ± 0.90	17.86 ± 0.87	10.37 ± 0.23
	pDDPM [21]	IXI	None	55.93 ± 0.28	21.79 ± 0.40	14.59 ± 0.47
	cDDPM [22]	IXI	None	**58.55 ± 0.78**	**24.74 ± 1.15**	11.94 ± 0.52
II. Self-Supervised	PII [119]	None	PII	30.38 ± 2.46	9.81 ± 1.93	23.44 ± 1.61
	DRAEM-Net [141]	None	DRAEM	24.78 ± 4.21	12.65 ± 1.90	**79.77 ± 2.37**
	Unet	None	DRAEM	40.75 ± 3.30	16.91 ± 0.38	76.03 ± 1.21
	Unet_{res}	IXI	DRAEM	45.80 ± 3.22	18.44 ± 0.47	77.43 ± 1.16
	SADM	IXI	DRAEM	50.81 ± 0.57	23.82 ± 0.32	73.77 ± 2.50
	SADM_{res} [25]	IXI	DRAEM	**60.53 ± 0.54**	**27.78 ± 0.14**	76.72 ± 1.30
III. Semi Supervised	Unet	None	BRATS	64.81 ± 0.21	11.82 ± 0.60	**24.83 ± 1.10**
	Unet_{res}	IXI	BRATS	67.01 ± 0.70	17.33 ± 1.31	19.93 ± 2.40
	SADM	IXI	BRATS	69.01 ± 0.21	25.25 ± 0.58	14.93 ± 0.51
	SADM_{res} [25]	IXI	BRATS	**69.68 ± 0.48**	**26.77 ± 0.65**	17.11 ± 1.78
	Unet	None	ATLAS	35.13 ± 2.97	46.30 ± 0.72	**29.11 ± 1.02**
	Unet_{res}	IXI	ATLAS	36.82 ± 4.18	47.36 ± 0.80	22.07 ± 2.20
	SADM	IXI	ATLAS	58.52 ± 0.60	46.40 ± 0.17	16.10 ± 1.10
	SADM_{res} [25]	IXI	ATLAS	**58.85 ± 0.44**	**47.64 ± 1.40**	17.77 ± 1.82

performance of all models when evaluating data sets within the same domain. However, the segmentation performance of the Unet and Unet_{res} is poor for data sets containing pathologies not encountered during training. In contrast, SADM and SADM_{res} enhance the segmentation performance on in-domain data while maintaining or improving the performance of unsupervised cDDPMs for unseen pathologies. These results demonstrate the potential of our framework to enhance segmentation performance on in-domain data while maintaining or even improving the performance of unsupervised cDDPMs for unseen pathologies. A qualitative comparison of the unsupervised and supervised predictions, as well as the final anomaly scores for SADM_{res}, is provided in [25] and Chapter 8.6.

5 Discussion

Machine learning, particularly deep learning, has shown considerable potential in medical image analysis, providing tools to assist radiologists in diagnosing a wide range of diseases [153, 154, 155, 156]. Most of these approaches rely on supervised learning, which requires large, annotated data sets. However, besides the challenges of acquiring such data sets, supervised models are inherently limited to detecting pathologies present in the training data, making them less effective at detecting unexpected or previously unseen conditions. UAD offers an alternative by learning the distribution of healthy anatomy and identifying deviations as potential anomalies. Among the different UAD approaches, reconstruction-based methods using GMs are particularly suitable, as they inherently provide pixel-level predictions and can generalize to a wide range of anomalies. However, a key challenge remains to achieve high reconstruction quality for healthy regions while ensuring that anomalies are not reproduced. Without proper constraints, GMs may learn to replicate both normal and abnormal structures, compromising their ability to distinguish anomalies from healthy tissue. To address this challenge, it is essential to enhance the reconstruction capabilities of GMs while regularizing the reconstruction process. The regularization ensures that the model captures the underlying structure of the data without simply copying the input image. In the following section, we review existing approaches to reconstruction-based UAD, discuss their limitations, and position our contributions within this context.

5.1 Advancements in Generative Models

The focus of reconstruction-based UAD has predominantly been on 2D slice-wise processing, where models reconstruct individual slices independently rather than considering the full 3D structure of brain MRI scans. A comparative study conducted in 2020 indicated that AEs and, in particular, VAEs are promising GMs for UAD [62]. However, both suffer from blurry reconstructions, limiting the detection of subtle anomalies, particularly those that manifest as minor intensity deviations [138, 149, 144].
To mitigate this, researchers have explored hierarchical AEs with Laplacian pyramids to improve reconstruction fidelity [104], along with various latent space optimizations such as adversarial training and disentangled representations [105, 102, 157]. Enhancements in VAEs include context-encoding methods [109] and age-conditioned models [158], while their probabilistic nature has been leveraged for transfer learning [110] and uncertainty estimation [150, 151, 159]. Additionally, vector-quantized VAEs have been combined with density estimation [160, 161]. Alongside these developments, GAN-based approaches have been explored, offering sharper reconstructions [111, 112, 113, 4, 140].
While these approaches improved reconstruction quality and detection performance, they share a fundamental limitation: processing adjacent slices in isolation ignores potentially valuable 3D information, which hinders the GM to capture the spatial relationships across slices. Therefore, we hypothesized that incorporating 3D information into the re-

construction process would enhance the GM's capacity to capture underlying anatomical structures, improving reconstruction quality and segmentation performance.

3D VAEs and Spatial Erasing

In light of this, our work [20] investigated using 3D VAEs for UAD in brain MRI. As presented in Section 4.1, 3D VAEs exhibited superior performance compared to their 2D counterparts, demonstrating enhanced reconstruction accuracy (Figure 4.1) and improved segmentation performance (Table 4.1). Our results suggest that incorporating additional spatial dimensions enables the model to better capture the underlying structure of the data, enhancing reconstruction quality. This finding aligns with most studies investigating 3D models for UAD [107, 146, 137]. However, a recent study on VQVAEs with transformers found no consistent improvement across all data sets when transitioning to 3D models [162]. These results indicate that the advantages of 3D models may depend on the specific model architecture and the characteristics of the data. One potential explanation for why 3D models may not always lead to improvements is their increased number of trainable parameters, which can contribute to overfitting. To address this, we adapted spatial erasing strategies to 3D data to improve the generalization and to counteract overfitting. The results of our experiments show that spatial erasing can enhance the segmentation performance for both 2D and 3D VAEs. Additionally, the results in Figure 4.1 indicate that the spatial erasing has a regularizing effect. While the erasing strategies slightly reduce the reconstruction quality for healthy regions, they lead to a more pronounced degradation in unhealthy regions, which in turn increases the contrast in the anomaly maps, enhancing segmentation.
Our results demonstrate that 3D VAEs with spatial erasing can improve segmentation performance. However, the overall performance remains moderate, achieving maximum Dice scores of 28.8% and 15.6% for the BRATS 2019 and the ATLAS v1 data sets, respectively. While VAEs with a dense latent space effectively capture the overall distribution of the training data without replicating the input, the absence of spatial information in the latent representation constrains reconstruction quality, ultimately limiting segmentation performance. Therefore, a common approach in the literature has been to enhance the information flow between the input and its reconstruction within AEs or VAEs. Using a spatial latent space [111] has been demonstrated to enhance reconstruction quality. However, this approach may also result in the model replicating the input image, contrary to the UAD principle [62, 146]. To prevent the models from merely replicating the input image, regularization techniques such as dropout [103] or denoising tasks [115, 163] have been explored. Furthermore, DDPMs that directly embed a denoising objective have emerged as promising GM for UAD [116, 162].
In contrast to the complete replacement of image components with noise, as seen in previous studies [68, 109, 20], in DDPMs, the noise is gradually introduced to the input image. During training, a denoising network is tasked with estimating and removing the added noise. Although DDPMs were primarily designed for image synthesis, they have been demonstrated to be effective for UAD in brain MRI [116, 162]. Typically, the sampling process of DDPMs is adapted to the reconstruction task by starting with partially noised images rather than pure noise. Furthermore, the study of [116] showed that using simplex noise instead of Gaussian noise can improve the UAD performance. Our experiments, as detailed in Section 4.2, substantiate that DDPMs outperform AEs and VAEs in terms of reconstruction quality (Table 4.2) and segmentation performance

(Table 4.4) by a margin. Nevertheless, our findings suggest that with the baseline DDPM (AnoDDPM), reconstructions remain imperfect, as demonstrated in 4.3. These results indicate that the noise introduced during the forward process of DDPMs leads to a loss of information, which is required to reconstruct the input image accurately.

Patched Diffusion Models

To improve the reconstruction quality, we proposed a patching strategy for DDPMs. Instead of introducing noise to the entire input image, we implemented the noise process solely within a designated patch, leveraging the global context of the unaltered patch surroundings during the denoising process. As demonstrated in Section 4.2, pDDPMs exhibit superior performance in reconstruction quality compared to AnoDDPM (Table 4.2). This superior performance suggests that the additional context provided by the patch surroundings can be effectively used by the denoising Unet, resulting in enhanced coherence between the input and the reconstruction. Moreover, the supplementary investigation of varying noise levels in Figure 4.2 demonstrates that pDDPMs can maintain high segmentation performance at higher noise levels compared to AnoDDPM, which apply noise to the entire image. This improved robustness suggests that our patching strategy allows us to achieve a better balance between reconstruction quality and regularization, which ultimately enhances the anomaly detection performance for the BRATS and MSLUB data sets. However, the segmentation performance of pDDPMs is dependent on the patch size. While our initial values performed well across the evaluated data sets, the additional hyperparameter may complicate finding an optimal solution across all possible anomalies. Furthermore, limitations are seen in the increased compute time due to patching and potential artifacts due to the stitching of reconstructed patches. Similar approaches to our patching strategy have been explored in recent studies. For instance, [164] implemented the patching within the frequency domain. Additionally, [165] investigated a selective noising strategy. In this approach, a two-stage strategy was proposed, whereby noise is selectively applied to the input image in the second stage based on a coarse residual map generated by RA [138] in the initial stage. However, this strategy remains dependent on the noise level at the initial stage, and the necessary post-processing steps introduce additional hyperparameters. Additionally, as with the limitations of our proposed pDDPMs, the two-stage nature of the approach results in higher complexity and necessitates additional computational resources.

Context-Conditioned Diffusion Models

To overcome the limitations of multiple reconstruction stages and patching artifacts, we proposed an efficient context-conditioning strategy for DDPMs. Accordingly, we directly integrated the context of the input image into the denoising process. An image encoder was employed to derive the additional context, generating an abstract feature representation of the input image. This feature representation was used to adjust the coarse shape and intensity distribution of the reconstruction, facilitating the generation of reconstructions that adapt locally to the input characteristics. A crucial outcome of Table 4.3 is the indication that the supplementary context in cDDPMs does not facilitate the replication of unhealthy structures, resulting in an increased $l1$-ratio. With these advancements, our cDDPMs outperform state-of-the-art baselines, including AnoDDPM

and pDDPMs, in terms of segmentation performance (Table 4.4) and reconstruction quality (Table 4.3) across the majority of data sets, while maintaining computational efficiency. Moreover, in alignment with the results of [166] our findings suggest that employing an ensembled reconstruction derived from a range of noise levels can enhance the reconstruction quality and segmentation performance of cDDPMs, reducing the reliance on the noise level hyperparameter. Additionally, our conditioning mechanism demonstrates enhanced domain adaptation capabilities for real and simulated intensity profiles, as we show in [22].
While our work aimed to improve reconstruction by conditioning on an abstract image representation, other studies have focused on refining the reverse diffusion process directly using the input image. THOR [167], which selectively reconstructs abnormal regions by iteratively reintegrating presumably healthy anatomy during denoising, aimed to reduce false positives and improve localization. In parallel, other studies have focused on improving computational efficiency. The study of [168] proposed a latent Bernoulli diffusion model, which compresses images into a binary latent space, reducing memory and computational costs while maintaining performance.

Our findings illustrate the inherent trade-off between reconstruction quality and regularization when employing GMs for reconstruction-based UAD. Dense AEs and VAEs inherently provide robust regularization due to their bottleneck structures. However, this often results in a compromise in reconstruction quality. Conversely, using a spatial latent space without sufficient regularization may result in the model merely replicating the input image. DDPMs offer a promising approach to balance reconstruction quality and regularization. However, the trade-off is now represented by the noise level. Adding too much noise can result in the loss of image information, which ultimately limits reconstruction quality and segmentation performance. On the other hand, insufficient added noise can lead to the model copying the input image, which hinders effective anomaly detection.
Addressing our first research question, we found that incorporating additional context, whether through 3D modeling, patching strategies or feature conditioning, can improve reconstruction accuracy and anomaly detection performance.
Nevertheless, it is difficult to simultaneously enhance the reconstruction quality and anomaly detection performance, as the necessary regularization results in imperfect reconstructions. This results in the introduction of minor reconstruction artifacts, which can be misinterpreted as anomalies, leading to the generation of false positives in the final segmentation. Hence, in addition to enhancing the reconstruction quality of the GMs, the anomaly scoring mechanism represents a crucial component for optimizing the performance of reconstruction-based UAD methods in brain MRI. Therefore, in the following section, we discuss our findings and advancements in anomaly scoring mechanisms to further improve the performance of UAD in brain MRI.

5.2 Advancements in Anomaly Scoring

In the majority of reconstruction-based UAD methods, the anomaly score is calculated based on the pixel-wise difference between the input image and its reconstruction. The most commonly employed metrics for this purpose are the MAE or MSE. However, these metrics only consider the intensity values of individual pixels without accounting for the

relationships between neighboring pixels. Consequently, these metrics rather highlight anomalies that manifest as intensity differences than structural changes [142, 149]. Furthermore, small reconstruction errors or noise in the reconstructions can be over-penalized. In industrial defect detection, the SSIM, which additionally accounts for structural changes, has been demonstrated to outperform solely intensity-based scoring functions [81]. Subsequently, this has also been demonstrated for AEs and VAEs in brain MRI [101, 144].

Ensembled SSIM for Anomaly Scoring

Given the substantial advancements achieved by DDPMs, we investigated possible synergies of SSIM with the robust reconstruction capabilities of DDPMs. Our results presented in 4.3 indicate that using SSIM to capture contextual information at the local neighborhood level can enhance the detection of anomalies, resulting in a substantial performance improvement. However, a closer investigation indicated that the SSIM is sensitive to the window size that determines the size of the considered neighborhoods. A larger window size captures more global structural information but may lose fine-grained details. In comparison, a smaller window size preserves local details but may miss the broader context of the image. This trade-off presents a challenge for the UAD task, as this task necessitates the detection of any anomaly, irrespective of its dimensions. Therefore, a single window size may not capture the full spectrum of possible pathologies. To enhance generalization, we proposed an extension of SSIM-based anomaly scoring by using an exponentially weighted ensemble of SSIM scores. We adaptively combined different window sizes to become more robust to different anomaly sizes and shapes. Our experiments demonstrate that while the ensembled SSIM does not consistently outperform the SSIM with individually tuned window sizes, it offers a more robust anomaly scoring mechanism that reduces the necessity for hyperparameter tuning.
Other studies have proposed using a KLD-based anomaly score or the use of activation maps like GradCAM [169] to refine anomaly scoring [108, 170]. However, Lagogiannis et al. [144] reported that the activation map approach exhibited poor generalization across different data sets, underscoring the importance of generalizability, a key advantage of the SSIM-ens score proposed in our study.

Mahalanobis Distance for Anomaly Scoring

So far, utilizing the SSIM, the contextual information came from the spatial surroundings of the compared pixels. However, DDPMs are not limited to reconstructing a single image; they can also sample a distribution of multiple reconstructions for a given input, providing additional context. We hypothesized that considering the distribution of reconstructions may be valuable in distinguishing anomalies from reconstruction artifacts. In the field of UAD, this concept has been investigated by performing probabilistic sampling with VAEs [62] and dropout-based Monte Carlo sampling with AEs [103]. However, these studies primarily concentrated on the mean of the generated reconstructions, which did not exhibit enhanced performance. Other approaches have employed VAEs or AEs with uncertainty estimation to normalize anomaly maps by estimating pixel-wise variance [150, 151, 159]. Although improving segmentation performance, these methods did not account for covariance across pixels. However, the inter-pixel covariance can

provide meaningful information, including global relationships across pixels, such as symmetries. Consequently, our study aimed to focus on these inter-pixel dependencies and variations across multiple pseudo-healthy reconstructions and to incorporate them into the anomaly scoring process. To this end, we proposed a novel anomaly scoring mechanism emphasizing the context of multiple reconstructions for each input. Our approach was to employ the MHD to quantify the discrepancy between input pixels and the distribution of pixels observed in healthy reconstructions. The results of our experiments, as detailed in Table 4.5, demonstrate that refining the SSIM using the MHD can enhance the segmentation performance. This enhancement suggests that incorporating the normal variability of the reconstructions can result in more reliable anomaly scores. Furthermore, according to our results, solely focusing on the variance across reconstructions and neglecting the inter-pixel covariance did not result in enhanced segmentation performance compared to the simple averaging approach. It is worth noting that the MHD has been used for outlier detection in brain MRI scans. However, its typical application has been at the sample level within some aggregated feature space [171, 172]. Moreover, Saase et al. applied the MHD in the pixel space using a healthy data set as a reference distribution, suggesting that simple statistical methods can compete with deep learning models [173]. However, our additional results in [24] indicate that relying solely on general population-based distributions can result in a discrepancy between individual test cases and the reference distribution, leading to poor segmentation performance. Accordingly, by employing the MHD in the pixel space of a generated distribution, we can provide a more robust anomaly scoring mechanism that can adapt to the specific characteristics of the individual subject, effectively capturing outliers. Furthermore, our approach is applicable to any probabilistic GM that can generate multiple reconstructions, providing a flexible and effective anomaly scoring mechanism for UAD in brain MRI.

Supervised Anomaly Scoring

So far, our approaches have followed the assumption that we do not have access to any labeled data, a common assumption in most brain MRI studies. Nevertheless, in practice, a small amount of labeled data is often available, which could be employed to enhance the anomaly scoring mechanism. Other studies have explored the application of weakly supervised anomaly detection [174, 175, 176]. Here, DDPMs were trained to remove specific pathologies based on image-level labels, generating pseudo-healthy counterfactuals for precise anomaly scoring. However, the effectiveness of this approach is constrained by the availability of labeled pathologies. Additionally, if no real labels exist, synthetic training data can be employed to simulate known anomalies for anomaly detection. However, directly training a supervised Unet to segment synthetic anomalies can result in overfitting and may not generalize effectively to real-world data or unseen pathologies, as also evidenced by the findings in [159, 144, 25, 22]. Therefore, in DISYRE [177, 178], synthetic anomalies are integrated into the diffusion process using Cold Diffusion [179]. Instead of Gaussian noise, synthetic anomalies are progressively introduced during the forward process, and a Unet learns to remove them in the reverse process while preserving healthy structures. Similarly, the study [180] proposed a conditioning-based approach where synthetic anomalies are added to healthy images and used as conditioning inputs during denoising. While these methods have shown promise compared to AE-based approaches, they were not compared to DDPM-based methods. Furthermore, their

reliance on synthetic anomalies raises concerns about generalization, particularly as the models are optimized for removing specific anomaly types.
Therefore, our objective was to identify an effective strategy for using supervision while maintaining the generalization capabilities of unsupervised methods. Accordingly, we proposed a framework to combine reconstruction-based UAD methods with a supervised scoring mechanism, which can be trained on a limited amount of real pathologies or synthetic anomalies while still being able to detect unseen anomalies. The results in Section 4.4 indicate that integrating supervision into the unsupervised pipeline can substantially enhance UAD performance. Additionally, training a supervised network to localize anomalies based on the discrepancy between the input image and a pseudo-healthy reconstruction can improve the segmentation performance and generalization capabilities of the supervised model. These findings suggests that providing information about the expected appearance of a healthy brain MRI may help the supervised model generalize more effectively to previously unseen pathologies. However, these experiments were conducted on a small subset of labeled pathologies. Therefore, further studies are needed to determine whether the proposed framework can consistently improve segmentation tasks, particularly when the supervised models are trained with larger data sets.
A similar framework was proposed in [141], wherein a generator network was trained to directly remove synthetic anomalies. However, our results indicate that this approach can only remove the anomalies seen during training and does not generalize well to unseen, real pathologies. In contrast, our approach can leverage the generalization capabilities of fully unsupervised methods that learn to reconstruct healthy anatomy, as opposed to the removal of specific anomalies. Furthermore, in comparison to the weakly supervised approaches presented in [174, 175, 176], our approach is capable of generalizing to unseen pathologies and can be integrated with any reconstruction-based UAD method.

In summary, our results demonstrate the importance of the anomaly scoring mechanism in enhancing the performance of reconstruction-based UAD methods in brain MRI. By leveraging spatial relationships, variability across reconstructions, and supervision, we developed robust scoring mechanisms that improve segmentation performance across diverse data sets. Therefore, in response to our second research question, we conclude that integrating additional context into anomaly scoring can enhance the segmentation performance.

5.3 Limitations and Implications for further Research

While our proposed methods improve the state-of-the-art of UAD in brain MRI, several limitations remain, which are discussed in the following sections.

Data set Dependency

Data is imperative for a comprehensive evaluation and development of novel approaches. However, existing data sets present significant challenges, particularly when evaluating UAD approaches. Typically, publicly available data sets focus on specific pathologies, while other anomalies or imaging artifacts might be present in the data without being annotated. For instance, the study of [165] highlights that substantial hypo-intense imaging artifacts exist in the ATLAS data set that are not considered anomalies by the provided annotations. However, as these artifacts deviate from the healthy distribution,

they may be flagged as anomalies by UAD methods, leading to false positives when compared to the provided annotations. This bias can skew model evaluations and may underestimate the capabilities of UAD methods compared to supervised approaches optimized for specific pathologies. Moreover, most data sets only include annotations for pathological conditions, with changes in healthy anatomy, such as those caused by space-occupying lesions, being overlooked. These changes can also result in the generation of false positive results in the segmentation process, as a successful UAD system might interpret these changes as anomalies. Future work should focus on developing more comprehensive evaluation data sets that include a diverse range of pathologies, anomalies, anatomical variations, and imaging artifacts.
Furthermore, apart from evaluating the segmentation performance of reconstruction-based models, it is important to assess their reconstruction quality of healthy and abnormal regions separately. A GM that only generates blurry shapes of the brain anatomy can achieve a high segmentation performance for hyper-intense anomalies but will struggle to detect subtle or structural anomalies, as shown in [142]. Therefore, a comprehensive evaluation of the reconstruction quality is essential to guarantee that the GMs can accurately capture the underlying structure of the data.
Although UAD methods reduce the necessity for annotated pathologies, they remain dependent on the accuracy of the healthy labels present within the data sets used for training. It has been demonstrated that even minor labeling inconsistencies, which result in anomalous data within the training set, can dramatically impair the efficacy of UAD methods [181, 162]. It is, therefore, essential that data sets used for training UAD models are carefully curated to ensure that they accurately represent the desired normal distribution. Additionally, approaches that are robust to impure training data [181] or leverage completely unlabeled data sets [159, 182] are seen as promising directions to improve the generalization capabilities of UAD models.

Bias and Generalization

An important aspect of UAD is the ability to detect a wide range of anomalies, regardless of their size, shape or appearance. However, while our methods achieve Dice scores of up to 70.6% for tumor segmentation in the BRATS data set, their performance for the detection of smaller, subtler anomalies is limited, with Dice scores of 24.2% for ATLAS and 11.6% for WMH data sets, respectively. These results are consistent across the evaluated UAD methods and highlight that UAD methods currently are not independent of the target pathology, and it remains challenging to reliably detect all types of anomalies with equal precision, regardless of their size or shape. This dependency can be introduced through post-processing, where, for example, the window size of a median filter may affect detection performance. Additionally, the threshold used to binarize anomaly maps for segmentation is typically determined on a calibration set, limiting the generalizability of UAD methods.
Beyond post-processing and thresholding, the GMs themselves may introduce biases toward detecting certain types of anomalies. For instance, the performance of DDPMs is influenced by parameters like the noise level or the type of noise used during the forward process. While simplex noise consistently outperformed Gaussian noise in our experiments, its specific structures may bias the detection capabilities toward particular anomaly morphologies. Furthermore, the optimal noise level can vary across different pathologies, as demonstrated in Figure 4.2. Additionally, hyperparameters, such as the

patch size in pDDPMs, can introduce biases. Smaller patch sizes favor detecting smaller pathologies and vice versa, making the choice of the patch size data set-dependent. These dependencies and biases highlight a general challenge of UAD methods, as they may not generalize well across different data sets or pathologies. To address these issues, future research should focus on developing more robust GMs and post-processing strategies, for instance by utilizing adaptive ensembling strategies for multiple median filters with different kernel sizes or GMs with different noise types and levels. Threshold selection could be improved by basing it on percentiles of the healthy distribution [183, 170, 184], or by incorporating uncertainty estimates from probabilistic GMs to adapt thresholds to the characteristics of the data.

Reconstruction Quality

A key challenge of reconstruction-based UAD methods is the accurate reconstruction of healthy structures. While our methods demonstrate strong reconstruction accuracy, instances remain where pathological regions are not fully replaced with healthy structures. The pathology structure is occasionally recreated with different intensities, violating the underlying assumption of reconstruction-based UAD methods. Conversely, reconstruction errors can introduce artifacts that can be misinterpreted as anomalies. While sophisticated anomaly scoring strategies can mitigate these errors, these limitations highlight the ongoing need for refinement in the reconstruction process. Future work should focus on improving the capacity of GMs to consistently reconstruct healthy anatomy, minimizing both the replication of pathological structures and the introduction of reconstruction artifacts. Promising directions include reconstruction-based models in combination with density estimation [160, 162, 117] or strategies where the reconstruction process is dependent on an adaptive masking strategy [165, 167]. Furthermore, additional metrics are required to quantify the "healing process" and assess reconstruction for healthy and unhealthy regions separately [185].

Scalability and Computational Constraints

Our experiments demonstrated that 3D approaches, such as 3D VAEs, can improve reconstruction quality and segmentation performance compared to their 2D counterparts. However, extending this to DDPMs for full-resolution volumes poses practical challenges due to the substantial computational demands and memory requirements. Even for 3D VAEs, working on large-resolution volumes requires significant memory. Hence, current GPU hardware may not support training 3D DDPMs at high resolutions, making it necessary to explore patching strategies or alternative architectures. This limitation has resulted in the predominance of slice-wise 2D processing, which neglects the spatial relationships between slices and may potentially limit the detection performance for certain pathologies. To address these challenges, alternative approaches such as latent diffusion models [186], which reduce computational complexity while preserving critical spatial information, have been adapted for UAD in brain MRI [162, 187]. Additionally, using wavelet diffusion models [188] for UAD in brain MRI or adopting sequential modeling of 2D slices [137] to DDPMs could provide efficient alternatives for processing 3D data.
A similar computational constraint arises when applying the MHD for anomaly scoring. The calculation of the MHD is computationally expensive, as it requires the inversion

of the covariance matrix. Consequently, without modifications, the proposed MHD approach may not be applicable to high-resolution 3D data. A potential solution is to subsample the covariance matrix, as not all elements may be necessary for robust anomaly scoring. Additionally, the MHD could be approximated using a low-rank decomposition to reduce computational complexity.

Future Directions and Clinical Applications

Addressing the outlined limitations requires further advancements in model architectures, anomaly scoring mechanisms, and the development of comprehensive benchmark data sets specifically tailored to the evaluation of UAD methods. One promising direction is refining the reconstruction process, for instance, through adaptive regularization strategies or density estimation, to enhance the separation of normal and abnormal structures. Additionally, improving efficiency and incorporating information from labeled data sets in semi supervised settings could help translate UAD research into clinical practice.
With these advancements, UAD methods have the potential to become valuable tools in clinical workflows, enabling the detection of unexpected or previously unseen anomalies. While supervised models excel at narrowly defined tasks where sufficient annotated data is available, they remain fundamentally constrained by their training data and can not generalize beyond predefined pathologies. Therefore, UAD research should aim to complement supervised methods, extending to cases where labeled data is scarce or unavailable.
A particularly impactful application of UAD lies in screening tasks, where the goal is not to detect a specific condition but to identify a wide spectrum of findings, including incidental and unanticipated pathologies. Moreover, for large-scale population studies, such as the Hamburg City Health Study [189], UAD could significantly reduce manual workload by automatically identifying deviations from normative distributions. Beyond diagnostics, UAD could contribute to quality control in medical imaging pipelines [166], identifying inconsistencies or artifacts that may compromise downstream analyses if left undetected. Additionally, it may aid in the discovery of novel biomarkers [190]. Another promising avenue is its potential role in medical education. By generating pseudo-healthy counterparts of pathological images, UAD could offer students and early-career radiologists a patient-specific healthy atlas, allowing for direct comparison and enhancing the understanding of the underlying anatomy and pathology.

In summary, while UAD is not a replacement for supervised diagnostic models, in the future, it could provide a powerful complementary tool for medical image analysis. Integrating UAD in future diagnostic systems could move beyond rigid supervised frameworks toward general diagnostic support, enabling new possibilities for early diagnosis and patient care.

6 Conclusion

Supervised deep learning has shown remarkable performance in medical image analysis. However, the reliance on annotated data limits its applicability, particularly for detecting rare, unexpected or previously unseen pathologies. UAD provides a more generalizable alternative by learning a reference distribution of healthy anatomy and identifying deviations as anomalies, making it especially valuable for detecting unexpected or rare pathologies. This thesis focused on advancing reconstruction-based UAD methods for brain MRI by introducing novel GMs and anomaly scoring mechanisms.

A key challenge in reconstruction-based UAD is balancing reconstruction quality and regularization. VAEs tend toward over-regularization due to their dense latent space, producing blurry reconstructions and limiting segmentation performance. Incorporating 3D context and spatial erasing improved the VAE performance. However, the gains were moderate, and the blurry reconstructions remained. DDPMs offered a better balance, leveraging a spatial latent space in combination with a denoising objective to achieve high-quality reconstructions, outperforming VAEs by a margin. However, challenges remained, and we observed intensity shifts and artifacts in the reconstructions of DDPMs. To address these limitations, we developed patched DDPMs and context-conditioned DDPMs, which integrate additional information from the input image to enhance reconstruction quality and reduce artifacts. Furthermore, advanced anomaly scoring mechanisms, such as an ensemble-based SSIM and a Mahalanobis distance-based approach, improved the distinction between genuine anomalies and reconstruction artifacts. Lastly, our proposed hybrid framework, SADM, demonstrated the potential to combine the generalization of unsupervised methods with the precision of supervised techniques.

Despite these contributions, challenges remain. Existing methods are highly sensitive to data set quality. Even minor labeling inconsistencies can negatively impact performance. Furthermore, pathology-specific biases persist, and the computational demands of 3D models limit their scalability to high-resolution volumes. Beyond these methodological challenges, the lack of standardized evaluation data sets and metrics hinders cross-study comparisons, making it difficult to assess progress in the field.

Future research should focus on developing efficient methods that are robust to labeling inconsistencies and generalize across diverse pathologies. A promising direction is seen in the probabilistic nature of DDPMs, which allows for the efficient integration of density estimation in reconstruction-based approaches. Furthermore, hybrid models such as our SADM offer the potential to combine the strong performance of supervised methods with the generalization ability of unsupervised approaches. Finally, standardized data sets, benchmarks, and clinical studies tailored to UAD in brain MRI will be essential for developing reliable and clinically applicable solutions.

With our proposed approaches, we have contributed to UAD research in brain MRI, establishing a solid foundation for future studies and advancing the development of clinically applicable methods. These contributions mark important steps toward the clinical adoption of UAD.

7 Summary

Magnetic resonance imaging (MRI) is a non-invasive diagnostic tool that is valuable for detecting neurological disorders. However, interpreting brain MRI scans is time-consuming, requires extensive expertise, and can be prone to errors. While deep learning models can support radiologists in diagnosis, most methods rely on supervised learning, which demands large-scale labeled data sets. This reliance limits their ability to detect rare, unexpected, or previously unseen conditions.
Unsupervised Anomaly Detection (UAD) presents an alternative by learning a reference distribution of healthy brain anatomy and identifying deviations as anomalies. In reconstruction-based UAD, Generative Models (GMs) are trained to reconstruct healthy brain MRI scans, assuming abnormalities will not be accurately reproduced after training. Anomaly maps are then generated based on the discrepancy between the input image and reconstruction. A key challenge in this approach is generating accurate reconstructions that preserve essential features without simply replicating the input images.
This thesis proposes the use of additional context in the reconstruction process of GMs and introduces novel anomaly scoring mechanisms to enhance the segmentation performance. The proposed models include 3D variational autoencoders with spatial erasing to use the inherently available 3D information of brain MRI scans. Furthermore, patched and context-conditioned diffusion models are proposed to leverage additional contextual information of the input image. For anomaly scoring, ensembles of structural similarity metrics and the Mahalanobis distance are employed to refine anomaly scoring. Additionally, a framework is proposed that integrates supervised anomaly scoring into the UAD pipeline.
Experiments demonstrate that context-aware reconstruction and anomaly scoring improve the segmentation performance. Moreover, the proposed hybrid UAD framework effectively combines the generalization capabilities of unsupervised approaches with the precise predictions of supervised methods. Despite these advancements, challenges remain. Sensitivity to data set quality affects performance, high-resolution 3D models require extensive computational resources, and the lack of standardized evaluation benchmarks complicates cross-study comparisons. Future research should focus on efficient, scalable models that are robust to labeling errors. Moreover, leveraging the probabilistic nature of diffusion models and integrating UAD and supervised methods in hybrid frameworks is promising. Lastly, establishing standardized data sets and clinical evaluation benchmarks will be essential for further progress.
This thesis advances UAD in brain MRI by introducing improvements to GMs and novel anomaly scoring mechanisms. The proposed methods build a solid foundation for the research field of UAD and contribute toward the application of UAD systems in real-world medical applications.

8 Publications

8.1 Three-dimensional deep learning with spatial erasing for unsupervised anomaly segmentation in brain MRI [20]

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

ORIGINAL ARTICLE

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

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

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

Abstract

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

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

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

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

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

Introduction

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

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

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

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

Marcel Bengs
marcel.bengs@tuhh.de

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

2 jung diagnostics GmbH, Hamburg, Germany

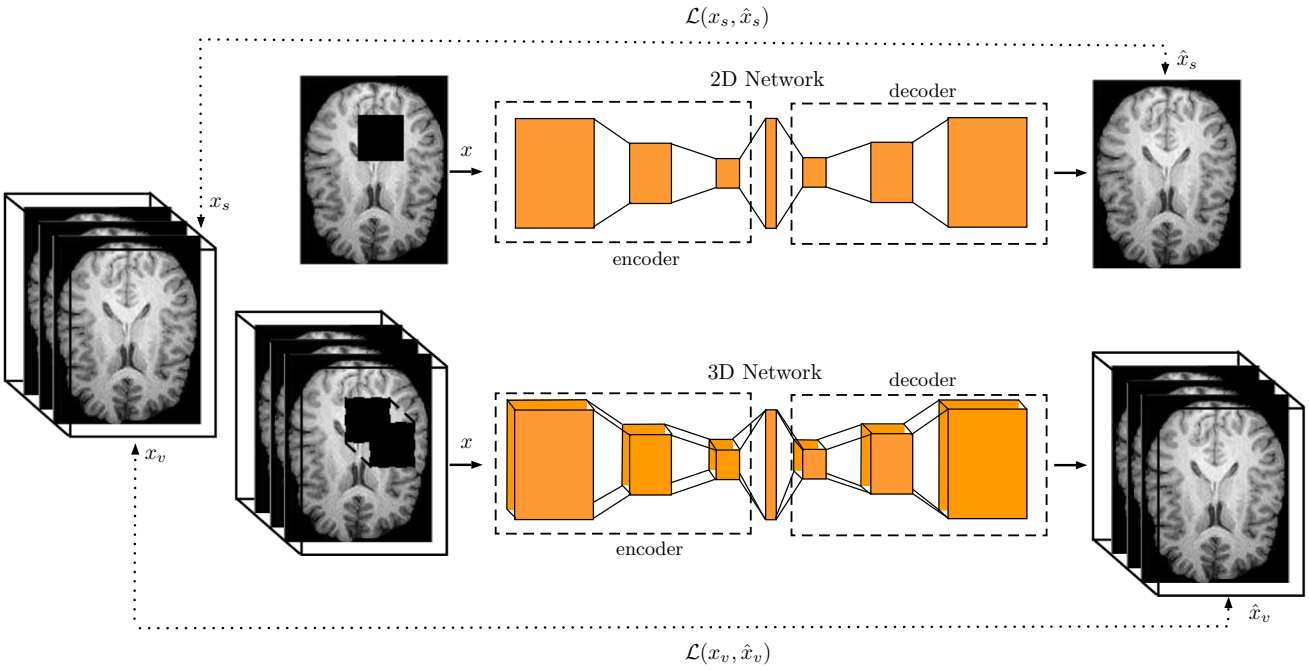


Fig. 1 Our approach for unsupervised anomaly segmentation using 3D deep learning combined with spatial input erasing. For the 2D network, only a single 2D slice x_s is used as input x and volumetric spatial context remains unexploited. Instead, our novel 3D approach receives an entire volume x_v as input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

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

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

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

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

Materials and methods

Data set

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

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

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

Deep learning methods

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

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

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

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

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

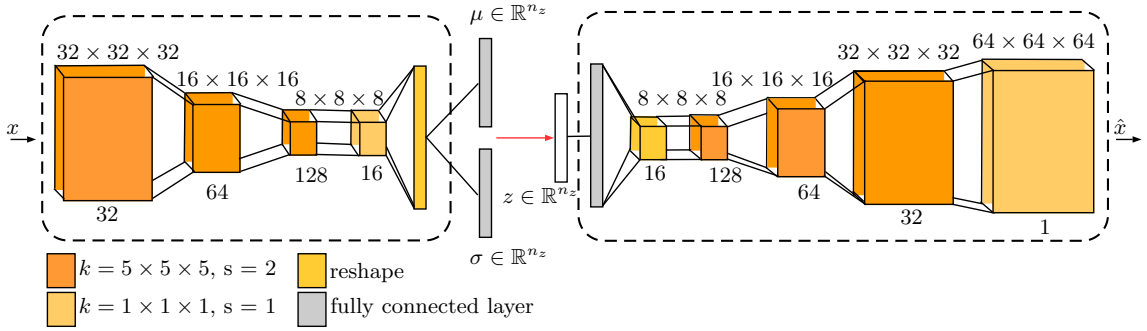


Fig. 2 Our backbone 3D VAD architecture receives input volume $x \in \mathbb{R}^{64\times64\times64}$ and encodes it to the lower-dimensional latent variable $z \in \mathbb{R}^{n_z}$, afterward the decoder reconstructs the output $\hat{x} \in \mathbb{R}^{64\times64\times64}$. The number over the boxes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

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

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

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

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

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

Training and evaluation

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

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

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

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

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

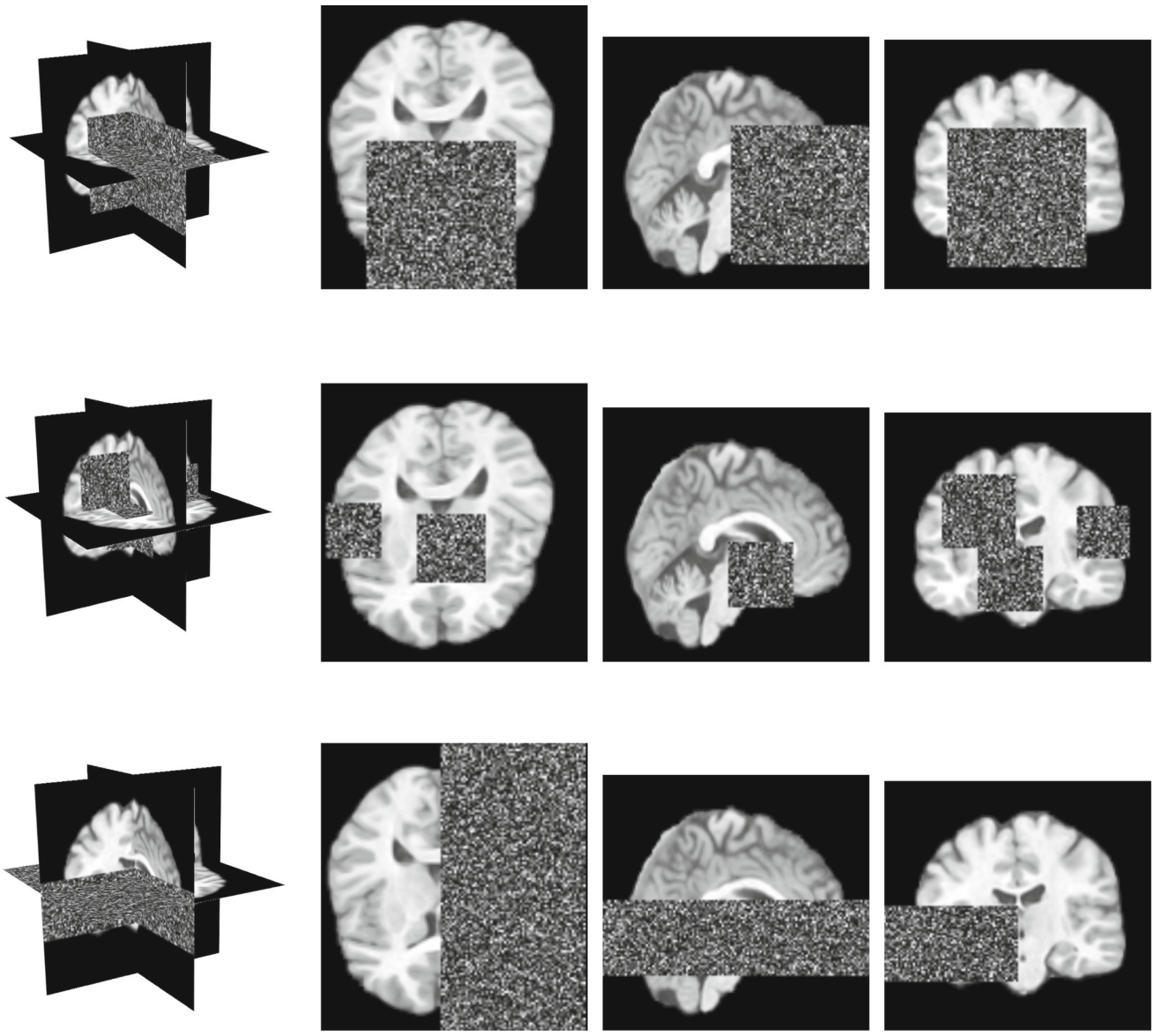


Fig. 3 Our 3D spatial input erasing methods. In each row, sectional planes of a volume with erasing are shown. Top row: We erase a single 3D cube with random location and size (Cube). Middle row: We erase multiple 3D cubes with random location and size (Multi-Cube). Bottom row: We erase an entire brain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

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

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

Results

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

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

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

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

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

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

All metrics are in percent

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

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

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

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

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

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

Discussion

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

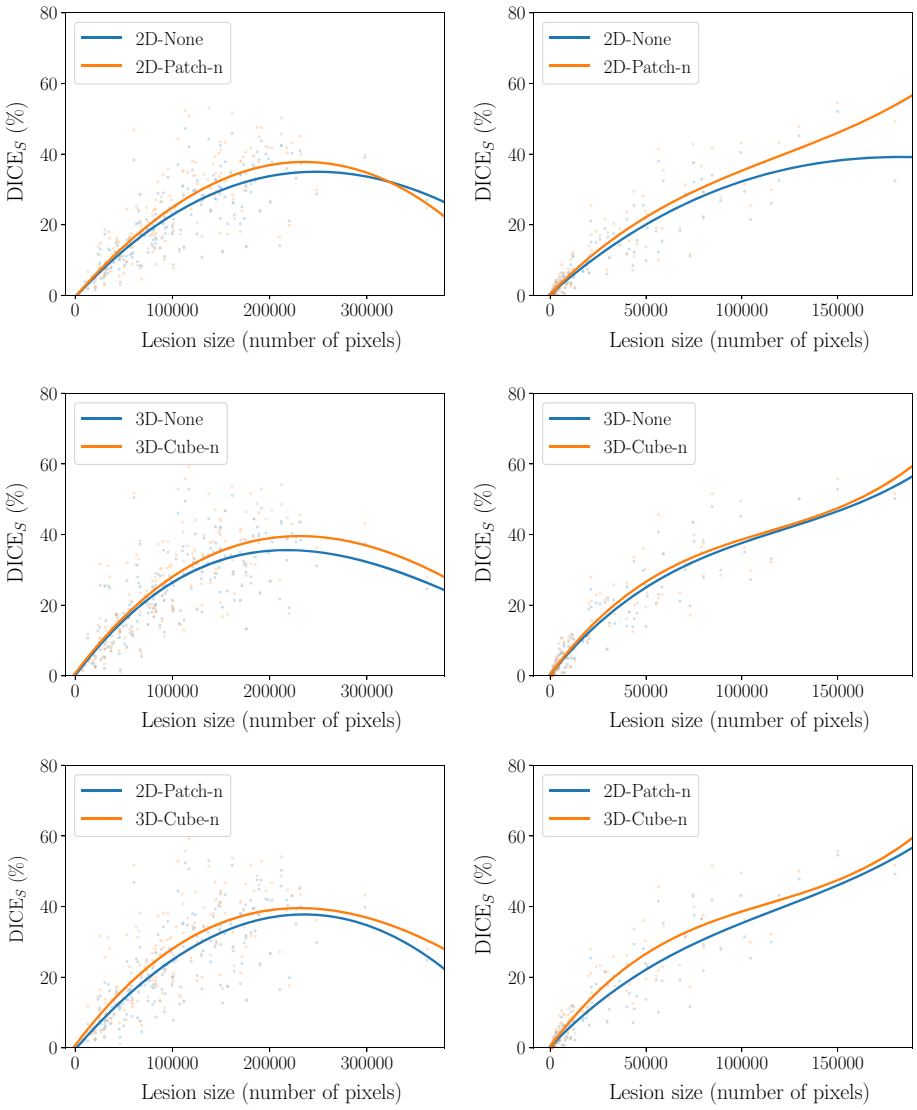


Fig. 4 Subject-wise DICE$_S$ over lesion size. Lesion size refers to the number of annotated pixels for the lesion. Results for the BraTS 2019 data set and ATLAS data set are shown left and right, respectively. (Top) Comparing 2D VAE with and without erasing; (Middle) Comparing 3D VAE with and without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

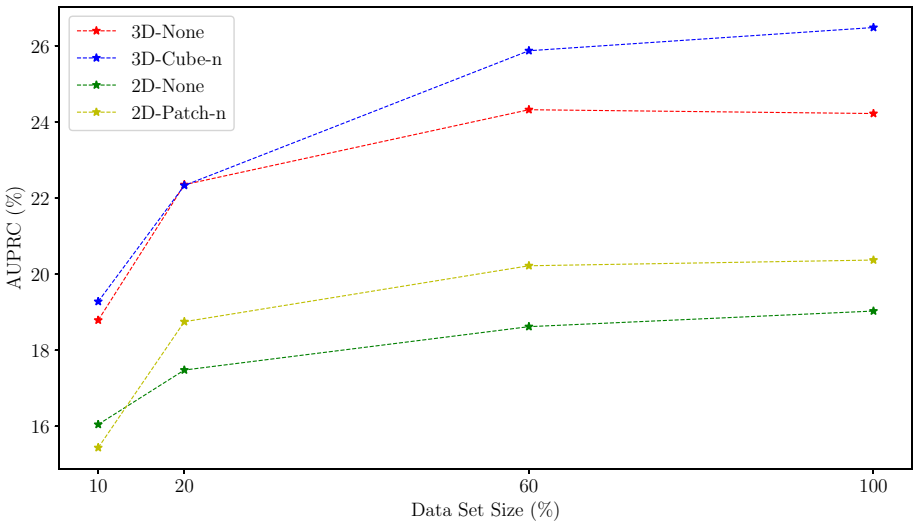


Fig. 5 Impact of data set size on the UAD performance. We train our methods with 10%, 20%, 60%, and 100% of the training data, shown is the average AUPRC using our two test data sets (BraTS 2019, ATLAS)

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

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

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

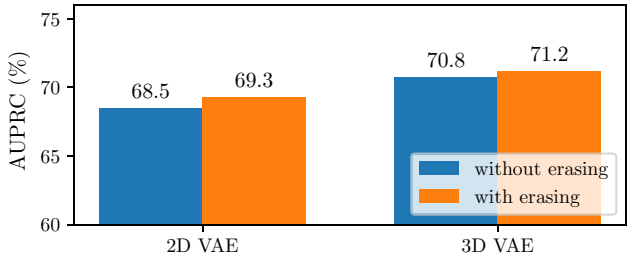


Fig. 6 Slice-wise anomaly detection for our baseline and best performing methods. Shown is the AUPRC on the combination of our test sets (BraTS 2019, ATLAS). 2D VAE with and without erasing refers to 2D-None and 2D-Patch-n, respectively. 3D VAE and without erasing refers to 3D-None and 3D-Cube-n, respectively

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

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

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

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

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

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

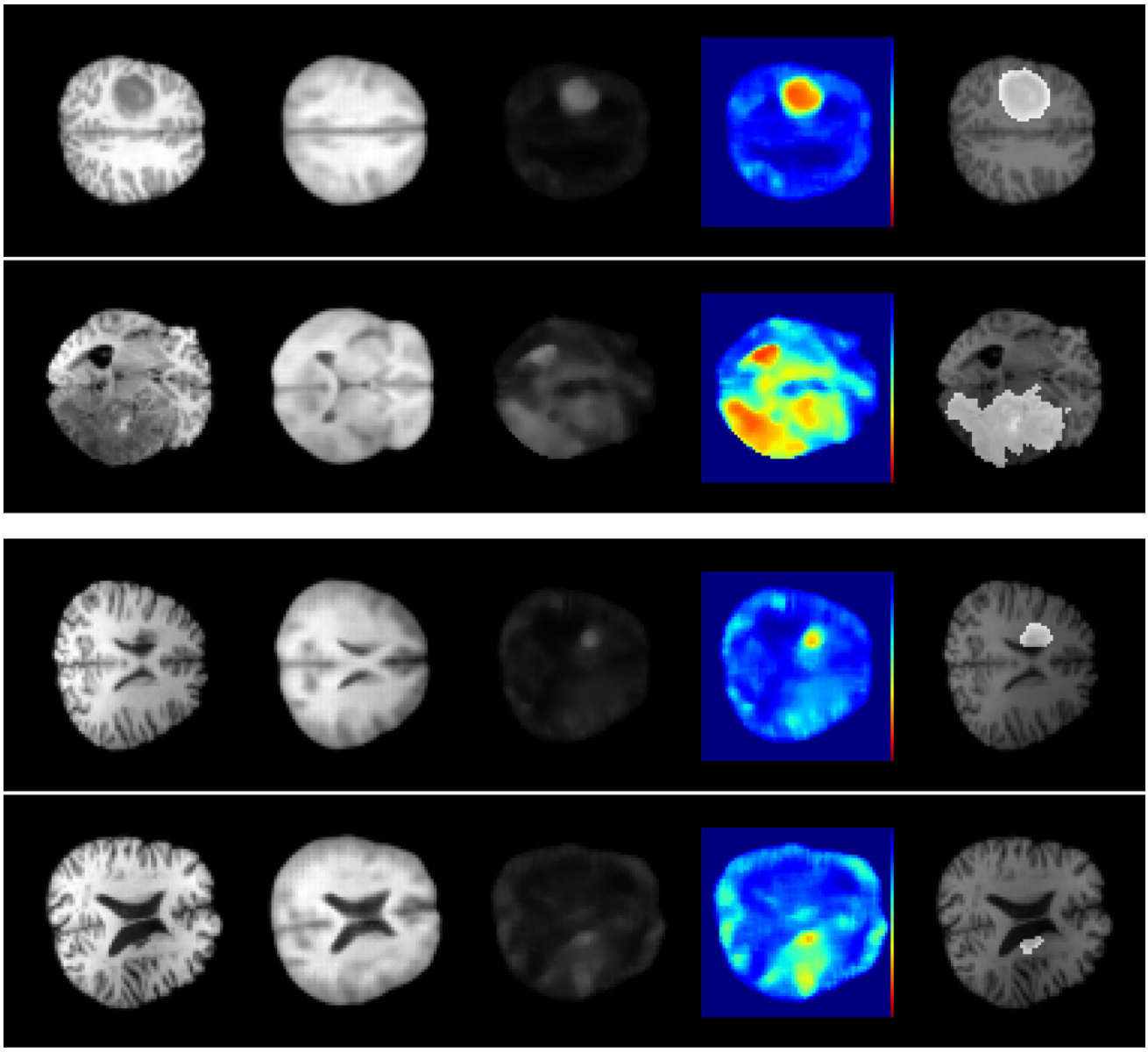


Fig. 7 Four example test cases using our best performing method 3D-Cube-n. From left to right: Input image, output image, difference image, heat-map difference image, and ground truth segmentation. The first two lines contain examples from the BraTS 2019 data set and the two bottom lines contain examples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

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

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

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

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

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

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

Conclusion

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

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

Declarations

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

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

References

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

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

8.2 Patched Diffusion Models for Unsupervised Anomaly Detection in Brain MRI [21]

Proceedings of Machine Learning Research 227:1019–1032, 2023 MIDL 2023 – Full paper track

Patched Diffusion Models for Unsupervised Anomaly Detection in Brain MRI

Finn Behrendt[1] FINN.BEHRENDT@TUHH.DE
Debayan Bhattacharya[1] DEBAYAN.BHATTACHARYA@TUHH.DE
Julia Krüger[2] JULIA.KRUEGER@JUNG-DIAGNOSTICS.DE
Roland Opfer[2] ROLAND.OPFER@JUNG-DIAGNOSTICS.DE
Alexander Schlaefer[1] SCHLAEFER@TUHH.DE
[1] *Institute of Medical Technology and Intelligent Systems, Hamburg University of Technology, Hamburg, Germany*

[2] *Jung Diagnostics GmbH, Hamburg, Germany*

Abstract

The use of supervised deep learning techniques to detect pathologies in brain MRI scans can be challenging due to the diversity of brain anatomy and the need for large, pixel-level annotated data sets. An alternative approach is to use unsupervised anomaly detection, which only requires sample-level labels of healthy brain anatomy to create a reference representation. This reference representation can then be compared to unhealthy brain anatomy in a pixel-wise manner to identify abnormalities. To accomplish this, generative models are needed to create anatomically consistent MRI scans of healthy brains. While recent diffusion models have shown promise in this task, accurately generating the complex structure of the human brain remains a challenge. In this paper, we propose a method that reformulates the generation task of diffusion models as a patch-based estimation of healthy brain anatomy, using spatial context to guide and improve reconstruction. We evaluate our approach on data of tumors and multiple sclerosis lesions and demonstrate a relative improvement of 25.1% in segmentation performance compared to existing baselines.

1. Introduction

Over the last decades, significant effort has been put into developing support tools that can assist radiologists in assessing medical images (Kawamoto et al., 2005). Convolutional neural networks (CNNs) have proven successful in this task due to their ability to process images effectively (Shen et al., 2017). However, supervised approaches that use CNNs have limitations, such as the need for large amounts of expert-annotated training data and the challenge of learning from noisy or imbalanced data (Ellis et al., 2022; Karimi et al., 2020; Johnson and Khoshgoftaar, 2019).

Unsupervised anomaly detection (UAD) is an alternative approach that can be trained with healthy samples only, eliminating the need for pixel-level annotations. During training, UAD models typically focus on reconstructing images from a healthy training distribution. When unseen, unhealthy anatomy is encountered at test time, high values in the pixel-wise reconstruction error indicate abnormalities.

Recently, denoising diffusion probabilistic models (DDPM) (Ho et al., 2020) have emerged as a state-of-the-art approach for image generation. As a result, they have also been applied to the problem of unsupervised anomaly detection (UAD) in brain MRI (Wyatt et al.,

2022; Pinaya et al., 2022a). DDPMs work by adding noise to an input image, then using a trained model to remove the noise and estimate or reconstruct the original image. Hence, in contrast to most autoencoder-based approaches, DDPMs preserve spatial information in their hidden representation of the input which is important for the image generation process (Rombach et al., 2022). However, applying noise to the entire image at once can make it difficult to accurately reconstruct the complex structure of the brain. To address this issue, we introduce patched DDPMs (pDDPMs) for UAD in brain MRI. In pDDPMs, we apply the forward diffusion process only on a small part of the input image and use the whole, partly noised image in the backward process to recover the noised patch. At test time, we use the trained pDDPM to sequentially noise and denoise a sliding patch within the input image and then stitch the individual denoised patches to reconstruct the entire image.
We evaluate our method on the public BraTS21 and MSLUB data sets and show that it significantly ($p < 0.05$) improves the tumor segmentation performance.

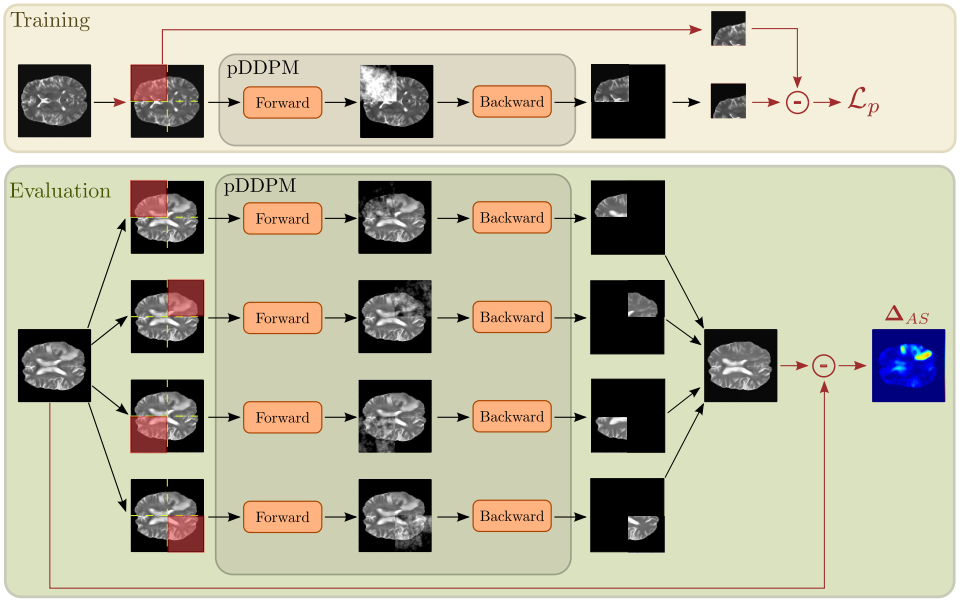


Figure 1: Schematic drawing of our method. From left to right: A patch is sampled within the input image, noise is added to that patch in the forward process and removed in the backward process. During evaluation, we stitch all patches and calculate the pixel-wise error as anomaly map $\mathbf{\Delta}_{AS}$.

2. Recent Work

In recent research on UAD in brain MRI, various architectures have been examined. Autoencoders (AE) and variational autoencoders (VAE) have demonstrated reliable training and fast inference, but their blurry reconstructions have hindered their effectiveness in UAD, as noted in (Baur et al., 2021). Therefore, research often focuses on understanding the image context better by adding spatial latent dimensions (Baur et al., 2018), multi-

resolution (Baur et al., 2020b), skip connections together with dropout (Baur et al., 2020a), or a denoising task as regularization (Kascenas et al., 2022). Similarly, modifications to VAEs aim to enforce the use of spatial context by spatial erasing (Zimmerer et al., 2019) or utilizing 3D information (Bengs et al., 2021; Behrendt et al., 2022). Other approaches propose restoration methods (Chen et al., 2020), uncertainty estimation (Sato et al., 2019), adversarial autoencoders (Chen and Konukoglu, 2018) or the use of encoder activation maps (Silva-Rodríguez et al., 2022). Also, vector-quantized VAEs have been proposed (Pinaya et al., 2022b). As an alternative to AE-based architectures, generative adversarial networks (GANs) have been applied to the problem of UAD (Schlegl et al., 2019). However, the unstable training nature of GANs makes their application very challenging. Furthermore, GANs suffer from mode collapse and often fail to preserve anatomical coherence (Baur et al., 2021). To alleviate this, inpainting approaches have been proposed that use the generator to inpaint erased patches during training (Nguyen et al., 2021). Lately, DDPMs have shown to be a promising approach for the task of UAD in brain MRI as they have scalable and stable training properties while generating sharp images of high quality (Wolleb et al., 2022; Wyatt et al., 2022; Sanchez et al., 2022; Pinaya et al., 2022a). While these approaches aim to estimate the entire brain anatomy at once, patch-based DDPMs have been proposed for image restoration (Özdenizci and Legenstein, 2023) and image inpainting (Lugmayr et al., 2022) in the domain of generic images. Patch-based DDPMs are a promising approach also for brain MRI reconstruction, as global context information about individual brain structure and appearance could be incorporated while estimating individual patches. However, current patch-based approaches either neglect the surrounding context of each patch (Özdenizci and Legenstein, 2023) or reconstruct patches from a fully noised image, which also impacts the surrounding context (Lugmayr et al., 2022). Thus, it is of interest to develop patch-based DDPMs that consider both the individual patch and its unperturbed surrounding context for the task of UAD in brain MRI.

3. Method

We apply the diffusion process of DDPMs in a patch-wise fashion, meaning that given the input image $\boldsymbol{x} \in \mathbb{R}^{C,W,H}$ with C channels, width W and height H, we add noise to a patch $\boldsymbol{p_k} \in \mathbb{R}^{C,h,w}$ with $h < H, w < W$ and $k = [1, ..., K]$. Subsequently, we reconstruct the patch to achieve a local estimate of the brain anatomy. Hereby, our motivation is a better understanding of image context by denoising image patches based on their unperturbed surrounding. Furthermore, we hypothesize that this would also lead to better anatomical coherence in the overall reconstruction of individual brains. As at test time anomalies can appear anywhere in the brain, we need to add and remove noise to the whole brain anatomy with our patch-wise approach. Therefore, we use a sliding window approach where we subsequently add noise to and remove noise from individual patches at positions that are evenly spaced across the image. Having covered the entire input image, we stitch all individual patch reconstructions into one image. This strategy allows estimating each local region in the input by using the spatial context of its surrounding which is assumed to be particularly helpful if the patch covers an anomaly. Our approach is shown in Figure 1.

3.1. DDPMs

In DDPMs, first, the image structure is gradually destroyed by noise and subsequently, the reverse denoising process is learned. During the forward process, adding noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to $\boldsymbol{x}_0$ follows a predefined schedule $\beta_1, ..., \beta_T$:

$$\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}), \text{with } \bar{\alpha}_t = \prod_{s=0}^{t}(1-\beta_t). \tag{1}$$

The time step t is sampled from $t \sim Uniform(1, ..., T)$ and controls how much noise is added to $\boldsymbol{x}_0$. For $t = T$ the image is replaced by pure Gaussian noise $\boldsymbol{x}_t = \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and for $t = 0$, $\boldsymbol{x}_t$ becomes $\boldsymbol{x}_0$.
In the backward process, the goal is to reverse the forward process and to recover $\boldsymbol{x}_0$.

$$\boldsymbol{x}_0 \sim p_\theta(\boldsymbol{x}_t) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t), \text{with } p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)). \tag{2}$$

$\boldsymbol{\mu}_\theta$ and $\boldsymbol{\Sigma}_\theta$ are estimated by a neural network with parameters θ. Following (Ho et al., 2020), we use an Unet (Ronneberger et al., 2015) for this task and keep $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t) = \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t\mathbf{I}$ fixed. To derive a tractable loss function, the variational lower bound (VLB) is used. By applying reformulations, Bayes rule and by conditioning $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ on $\boldsymbol{x}_0$, minimizing the VLB can be approximated by the simpler loss derivation $\mathcal{L}_{simple} = ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)||^2$. In this work, we utilize the $l1$-error and change the objective to directly estimate $\boldsymbol{x}_0^{rec} \sim p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_t, t)$, leading to $\mathcal{L}_{rec} = |\boldsymbol{x}_0 - \boldsymbol{x}_0^{rec}|$. For sampling images with DDPMs, typically step-wise denoising is applied for all time steps starting from $t = T$. As this comes at the cost of long sampling times, in this work we directly estimate $\boldsymbol{x}_0^{rec} \sim p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_t, t)$ at a fixed time step t_{test}. This simplification is possible since we do not aim to generate new images from noise but are interested in reconstructing a given image.

3.2. Patched DDPMs

As aforementioned, with patched DDPMs, we apply the forward and backward process in a patched fashion. During training, we sample the patches either at random positions or from a fixed grid defined as follows. We partition $\boldsymbol{x}$ into K patch regions that are evenly spaced across $\boldsymbol{x}$. The number of possible patch regions in $\boldsymbol{x}$ is derived as $K = \lceil \frac{W-w}{w} \rceil + \lceil \frac{H-h}{h} \rceil + 2$, where $\lceil . \rceil$ denotes the ceiling operation. From this grid, we uniformly sample an index k.
During the forward step of the diffusion process, i.e., the noising step we sample the noised image $\boldsymbol{x}_t$ only at the given patch position $\boldsymbol{p_k}$. Consider $\boldsymbol{M_p} \in \mathbb{R}^{C,H,W}$ a binary mask where the pixels that overlap with $\boldsymbol{p_k}$ are set to one and pixels that do not overlap with $\boldsymbol{p_k}$ are set to zero. We obtain the partly noised image as

$$\tilde{\boldsymbol{x}}_{\boldsymbol{t}} = \boldsymbol{x}_t \odot \boldsymbol{M_p} + \boldsymbol{x}_0 \odot \neg \boldsymbol{M_p} \tag{3}$$

where $\odot$ denotes element-wise multiplication. In the backward process, $\tilde{\boldsymbol{x}}_{\boldsymbol{t}}$ is fed to the denoising network to estimate the given noise area. The denoised image is derived as $\tilde{\boldsymbol{x}}_0^{rec} \sim p_\theta(\boldsymbol{x}_0|\tilde{\boldsymbol{x}}_{\boldsymbol{t}}, t)$. To train the patch-wise denoising task, we optionally use an objective function $\mathcal{L}_p$ adapted from $\mathcal{L}_{rec}$, where we calculate $\mathcal{L}_p = |(\boldsymbol{x}_0 - \tilde{\boldsymbol{x}}_0^{rec}) \odot \boldsymbol{M_p}|$ based on the noised region within $\boldsymbol{p_k}$, ignoring the surrounding area.
During Evaluation, for every $k \in [0, ..., K]$, we subsequently perform the diffusion process

based on the patch $\boldsymbol{p_k}$. After the reconstruction of all patch regions, we use the reconstructed patches $[\boldsymbol{p}_0^{rec}, ..., \boldsymbol{p}_K^{rec}]$ and stitch them with respect to their original position in the input image to retain the full reconstruction of $\boldsymbol{x}_0$. In the case of overlapping patches, we average the overlapping regions of the reconstructed patches.

4. Experimental setup

4.1. Data

We use the publicly available IXI data set as healthy reference distribution for training. The IXI data set consists of 560 pairs of T1 and T2-weighted brain MRI scans, acquired in three different hospital sites. From the training data, we use 158 samples for testing and partition the remaining data set into 5 folds of 358 training samples and 44 validation samples for cross-validation and stratify the sampling by the age of the patients.
For evaluation, we utilize two publicly available data sets, namely the Multimodal Brain Tumor Segmentation Challenge 2021 (BraTS21) data set (Baid et al., 2021; Bakas et al., 2017; Menze et al., 2014), and the multiple sclerosis data set from the University Hospital of Ljubljana (MSLUB) (Lesjak et al., 2018).
The BraTS21 data set consists of 1251 brain MRI scans of four different weightings (T1, T1-CE, T2, FLAIR). We split the data set into an unhealthy validation set of 100 samples and an unhealthy test set of 1151 samples. The MSLUB data set consists of brain MRI scans of 30 patients with multiple sclerosis (MS). For each patient T1, T2, and FLAIR-weighted scans are available. We split the data into an unhealthy validation set of 10 samples and an unhealthy test set of 20 samples. For both evaluation data sets, expert annotations in form of pixel-wise segmentation maps are available.
Across our experiments, we utilize T2-weighted images from all data sets. To align all MRI scans we register the brain scans to the SRI24-Atlas (Rohlfing et al., 2010) by affine transformations. Next, we apply skull stripping with HD-BET (Isensee et al., 2019). Note that these steps are already applied to the BraTS21 data set by default. Subsequently, we remove black borders, leading to a fixed resolution of $[192 \times 192 \times 160]$ voxels. Lastly, we perform a bias field correction. To save computational resources, we reduce the volume resolution by a factor of two resulting in $[96 \times 96 \times 80]$ voxels and remove 15 top and bottom slices parallel to the transverse plane.

4.2. Implementation Details

We evaluate our proposed method *pDDPM*, against multiple established baselines for UAD in brain MRI. These include *AE*, *VAE* (Baur et al., 2021), their sequential extension *SVAE* (Behrendt et al., 2022), and denoising AEs *DAE* (Kascenas et al., 2022). We also compare simple thresholding *Thresh* (Meissen et al., 2022), and the GAN-based *f-AnoGAN* (Schlegl et al., 2019). Additionally, we chose *DDPM* (Wyatt et al., 2022) as a counterpart to our proposed method. We implement all baselines based on their original publications with the following individual adaptations that have been shown to improve training stability and performance. For *VAE* and *SVAE*, we set the value of β_{VAE} to 0.001. For *f-AnoGAN*, we set the latent size to 128 and the learning rate to $1e-4$.
For *DDPM* and *pDDPM*, we utilize structured simplex noise, rather than Gaussian noise,

as it is known to better capture the natural frequency distribution of MRI images (Wyatt et al., 2022). For training, we uniformly sample $t \in [1, T]$ with $T = 1000$, and at test time, we choose a fixed value of $t_{test} = \frac{T}{2} = 500$. We choose a linear schedule for β_t, ranging from $1e-4$ to $2e-2$ and use an Unet similar to (Dhariwal and Nichol, 2021) as a denoising network. For each channel dimension $C_f \in [128, 128, 256]$, the Unet consists of a stack of 3 residual layers and downsampling convolutions. This structure is mirrored in the upsampling path with transposed convolutions. Skip connections connect the layers at each resolution. In each residual block, groupnorm is used for normalization and SiLU (Elfwing et al., 2018) acts as activation function before convolution. For time step conditioning, the time step is first encoded using a sinusoidal position embedding and then projected to a vector that matches the channel dimension. This is added to the feature representation using scale-shift-norm (Perez et al., 2018) in each residual block. Unless specified otherwise, all models are trained for a maximum of 1600 epochs, and the best model checkpoint, as determined by performance on the healthy validation set, is used for testing. We process the volumes in a slice-wise fashion, uniformly sampling slices with replacement during training and iterating over all slices to reconstruct the full volume at test time. The models were trained on NVIDIA V100 GPUs (32GB) using Adam as the optimizer, a learning rate of $1e-5$, and a batch size of 32. The code for this work is available at `https://github.com/FinnBehrendt/patched-Diffusion-Models-UAD`.

4.3. Post-Processing and Anomaly Scoring

During training, all models aim to minimize the $l1$ error between the input and its reconstruction. At test time, we use the reconstruction error as a pixel-wise anomaly score $\boldsymbol{\Delta}_{AS} = |\boldsymbol{x}_0 - \boldsymbol{x}_0^{rec}|$, where high values indicate larger reconstruction errors and vice versa. Given the hypothesis that the models will fail to reconstruct unhealthy brain anatomy, we assume that anomalies are located at regions of high reconstruction errors. We apply several post-processing steps that are commonly used in the literature (Baur et al., 2021; Zimmerer et al., 2019). Before binarizing $\boldsymbol{\Delta}_{AS}$, we use a median filter with kernel size $K_M = 5$ to smooth $\boldsymbol{\Delta}_{AS}$ and perform brain mask eroding for 3 iterations. Having binarized $\boldsymbol{\Delta}_{AS}$, we apply a connected component analysis, removing segments with less than 7 voxels. To achieve a threshold for binarizing $\boldsymbol{\Delta}_{AS}$, we perform a greedy search based on the unhealthy validation set where the threshold is determined by iteratively calculating Dice scores for different thresholds. The best threshold is then used to calculate the average Dice score on the unhealthy test set (DICE). Furthermore, we report the average Area Under Precision-Recall Curve (AUPRC) and report the mean absolute reconstruction error ($l1$) of the test split from our healthy IXI data set.

4.4. Statistical Testing

For significance tests, we employ a permutation test from the MLXtend library (Raschka, 2018) with a significance level of $\alpha = 5\%$ and 10,000 rounds of permutations. The test calculates the two models' mean difference of the Dice scores for each permutation. The resulting p-value is determined by counting the number of times the mean differences were equal to or greater than the sample differences, divided by the total number of permutations.

Table 1: Comparison of the evaluated models with the best results highlighted in bold. *fixed sampling* denotes that patch positions are sampled from a fixed grid, in contrast to *random sampling*, where patch positions are randomly sampled. $\mathcal{L}_p$ denotes calculating the reconstruction loss only on the patch region whereas $\mathcal{L}_{rec}$ denotes calculating the reconstruction loss for the whole image. For all metrics, mean ± standard deviation across the different folds are reported.

	BraTS21		MSLUB		IXI
Model	**DICE** [%]	**AUPRC** [%]	**DICE** [%]	**AUPRC** [%]	$l1$ $(1e-3)$
Thresh (Meissen et al., 2022)	19.69	20.27	6.21	4.23	145.12
AE (Baur et al., 2021)	32.87±1.25	31.07±1.75	7.10±0.68	5.58±0.26	30.55±0.27
VAE (Baur et al., 2021)	31.11±1.50	28.80±1.92	6.89±0.09	5.00±0.40	31.28±0.71
SVAE (Behrendt et al., 2022)	33.32±0.14	33.14±0.20	5.76±0.44	5.04±0.13	28.08±0.02
DAE (Kascenas et al., 2022)	37.05±1.42	44.99±1.72	3.56±0.91	5.35±0.45	**10.12±0.26**
f-AnoGAN (Schlegl et al., 2019)	24.16±2.94	22.05±3.05	4.18±1.18	4.01±0.90	45.30±2.98
DDPM (Wyatt et al., 2022)	40.67±1.21	49.78±1.02	6.42±1.60	7.44±0.52	13.46±0.65
pDDPM + random sampling + $\mathcal{L}_{rec}$	44.47±2.34	48.84±2.71	9.41±0.96	9.13±1.13	14.08±0.77
pDDPM + fixed sampling + $\mathcal{L}_{rec}$	47.81±1.15	52.38±1.17	**10.47±1.27**	**10.58±0.85**	12.12±0.76
pDDPM + fixed sampling + $\mathcal{L}_p$	**49.00±0.84**	**54.07±1.06**	10.35±0.69	9.79±0.4	11.05±0.15

5. Results

Unless stated otherwise, for *pDDPM*, we use patch dimensions of $h = w = \frac{H}{2} = \frac{W}{2} = 48$. The comparison of our pDDPM with the baseline models is shown in Table 1. Like *DAE*, the *DDPM* shows relatively high performance on the BraTS21 data set, but its performance on the MSLUB data set is moderate. In contrast, our *pDDPM* outperforms all baselines on both data sets regarding DICE and AUPRC, with statistical significance for the BraTS21 data set ($p < 0.05$). Considering the reconstruction quality by means of $l1$ error on healthy data, the *DAE* shows the lowest reconstruction error, followed by *pDDPM*.
Qualitatively, we observe smaller reconstruction errors from *pDDPMs* compared to *DDPMs* for healthy brain anatomy as shown in Figure 2. Examples of reconstructions from other baseline models can be found in Appendix 4. As seen in Figure 3, a patch size of 60×60 pixels results in the best performance. Additionally, there is a peak in performance when the noise level at test time is $t_{test} = 400$. A visualization of different noise levels is provided in Appendix B and ablation studies for the MSLUB data set are available in Appendix C.

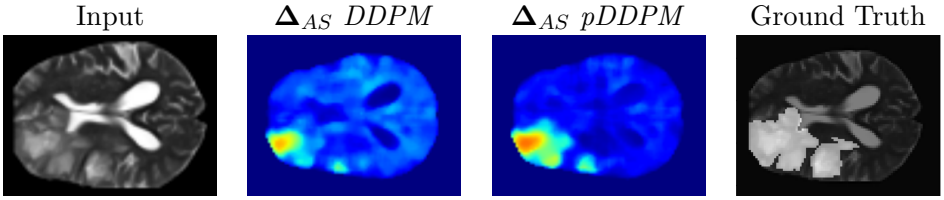


Figure 2: Visualization of input, errormap and the ground truth for *DDPM* and *pDDPM* for the Brats21 data set.

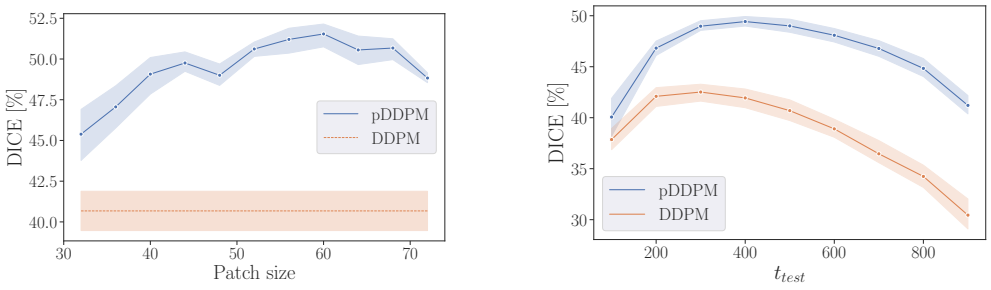


Figure 3: DICE for different patch sizes (left) and noise levels at test time t_{test} (right) for the BraTS21 data set. We report the average DICE across the 5 cross-validation folds. Standard deviations are visualized as enveloping intervals.

6. Discussion & Conclusion

Our approach frames the reconstruction of healthy brain anatomy as patch-based denoising, allowing to incorporate context information about individual brain structure and appearance when estimating brain anatomy. We show that *pDDPMs* outperform both their non-patched counterparts and various baseline methods with significant differences for the BraTS21 data set ($p < 0.05$).

Our results indicate that the image context around the noised patch can be used effectively by the model to replace potential anomalies covered by noise patches with estimates of healthy anatomy. From the performance improvements resulting from selecting patches from fixed positions and minimizing $\mathcal{L}_p$ rather than $\mathcal{L}_{rec}$, we conclude that it is helpful to focus on pre-defined local patches during training. By stitching the individual patches, we achieve sharp reconstructions without the downside of reconstructing too much unhealthy anatomy. Note that this trade-off is influenced by both, the noise level t_{test} and the patch size as shown in Figure 3. While our initial values for these hyper-parameters already show robust performance improvements across both data sets, further tuning results in more optimal settings for certain anomalies. To enhance generalization across different anomalies, employing an ensemble of different patch sizes and noise levels, as demonstrated in (Graham et al., 2022), is a promising direction for future research. Evaluating the reconstruction quality by means of $l1$ error, *DAE* shows superior results to *pDDPM*. However, *DAE* is able to reconstruct unhealthy anatomy which increases false negative predictions and thus decreases the UAD performance. We observe that accurately identifying MS lesions in T2-weighted MRI scans is challenging, and the limited number of samples makes it hard to achieve statistically significant results. However, our *pDDPMs* show promising improvements on the MSLUB data set, suggesting that it could be useful to address the challenges of detecting MS lesions. To further improve the UAD performance, using FLAIR-weighted MRI scans or enriching the anomaly scoring by structural differences could be valuable.

Our proposed approach has shown promising results in terms of UAD performance, however, it does have the drawback of an increase in inference time. While parallel computing could alleviate the increase in inference time, future work could focus on guiding the denoising process by spatial context more efficiently.

Acknowledgments

This work was partially funded by grant number KK5208101KS0 and ZF4026303TS9 and by the Free and Hanseatic City of Hamburg (Interdisciplinary Graduate School) from University Medical Center Hamburg-Eppendorf

References

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

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(1):1–13, 2017.

Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *International MICCAI brainlesion workshop*, pages 161–169. Springer, 2018.

Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Bayesian skip-autoencoders for unsupervised hyperintense anomaly detection in high resolution brain mri. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1905–1909. IEEE, 2020a.

Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Scale-space autoencoders for unsupervised anomaly segmentation in brain mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 552–561. Springer, 2020b.

Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study. *Medical Image Analysis*, page 101952, 2021.

Finn Behrendt, Marcel Bengs, Debayan Bhattacharya, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Capturing inter-slice dependencies of 3d brain mri-scans for unsupervised anomaly detection. In *Medical Imaging with Deep Learning*, 2022.

Marcel Bengs, Finn Behrendt, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Three-dimensional deep learning with spatial erasing for unsupervised anomaly segmentation in brain mri. *International journal of computer assisted radiology and surgery*, 16(9): 1413–1423, 2021.

Xiaoran Chen and Ender Konukoglu. Unsupervised Detection of Lesions in Brain MRI using Constrained Adversarial Auto-encoders. In *International Conference on Medical Imaging with Deep Learning (MIDL)*, Proceedings of Machine Learning Research. PMLR, 2018.

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

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

Randall J. Ellis, Ryan M. Sander, and Alfonso Limon. Twelve key challenges in medical machine learning and solutions. *Intelligence-Based Medicine*, 6:100068, 2022. ISSN 2666-5212.

Mark S Graham, Walter HL Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. *arXiv preprint arXiv:2211.07740*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, et al. Automated brain extraction of multisequence mri using artificial neural networks. *Human brain mapping*, 40(17):4952–4964, 2019.

Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020.

Antanas Kascenas, Nicolas Pugeault, and Alison Q O'Neil. Denoising autoencoders for unsupervised anomaly detection in brain mri. In *International Conference on Medical Imaging with Deep Learning (MIDL)*, Proceedings of Machine Learning Research. PMLR, 2022.

Kensaku Kawamoto, Caitlin A Houlihan, E Andrew Balas, and David F Lobach. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *Bmj*, 330(7494):765, 2005.

Žiga Lesjak, Alfiia Galimzianova, Aleš Koren, Matej Lukin, Franjo Pernuš, Boštjan Likar, and Žiga Špiclin. A novel public mr image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. *Neuroinformatics*, 16(1):51–63, 2018.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

Felix Meissen, Georgios Kaissis, and Daniel Rueckert. Challenging current semi-supervised anomaly segmentation methods for brain mri. In *International MICCAI brainlesion workshop*, pages 63–74. Springer, 2022.

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

Bao Nguyen, Adam Feldman, Sarath Bethapudi, Andrew Jennings, and Chris G Willcocks. Unsupervised region-based anomaly detection in brain mri with adversarial image inpainting. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1127–1131. IEEE, 2021.

Ozan Özdenizci and Robert Legenstein. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Walter HL Pinaya, Mark S Graham, Robert Gray, Pedro F Da Costa, Petru-Daniel Tudosiu, Paul Wright, Yee H Mah, Andrew D MacKinnon, James T Teo, Rolf Jager, et al. Fast unsupervised brain anomaly detection and segmentation with diffusion models. *arXiv preprint arXiv:2206.03461*, 2022a.

Walter HL Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Unsupervised brain imaging 3d anomaly detection and segmentation with transformers. *Medical Image Analysis*, 79:102475, 2022b.

Sebastian Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*, 3(24), April 2018. doi: 10.21105/joss.00638. URL http://joss.theoj.org/papers/10.21105/joss.00638.

Torsten Rohlfing, Natalie M Zahr, Edith V Sullivan, and Adolf Pfefferbaum. The sri24 multichannel atlas of normal adult human brain structure. *Human brain mapping*, 31(5): 798–819, 2010.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Pedro Sanchez, Antanas Kascenas, Xiao Liu, Alison Q O'Neil, and Sotirios A Tsaftaris. What is healthy? generative counterfactual diffusion for lesion localization. In *MICCAI Workshop on Deep Generative Models*, pages 34–44. Springer, 2022.

Kazuki Sato, Kenta Hama, Takashi Matsubara, and Kuniaki Uehara. Predictable uncertainty-aware unsupervised deep anomaly segmentation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2019. doi: 10.1109/IJCNN.2019.8852144.

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.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221, 2017.

Julio Silva-Rodríguez, Valery Naranjo, and Jose Dolz. Constrained unsupervised anomaly segmentation. *Medical Image Analysis*, 80:102526, 2022.

Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. *arXiv preprint arXiv:2203.04306*, 2022.

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

David Zimmerer, Simon Kohl, Jens Petersen, Fabian Isensee, and Klaus Maier-Hein. Context-encoding variational autoencoder for unsupervised anomaly detection. In *International Conference on Medical Imaging with Deep Learning–Extended Abstract Track*, 2019.

Appendix A. Exemplary reconstructions for all Baselines

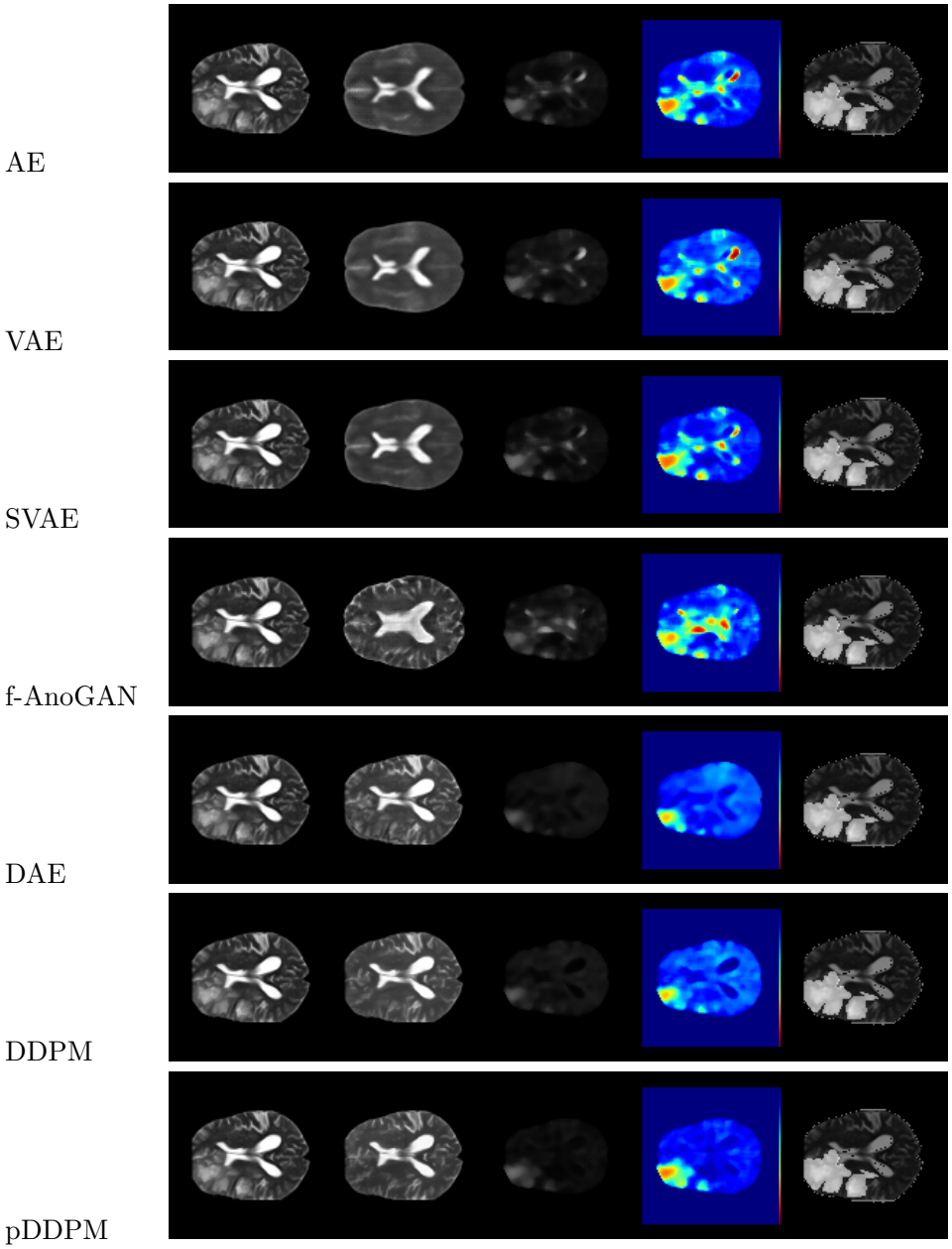


Figure 4: Qualitative evaluation of reconstructions from different models. From top to bottom: AE, VAE, SVAE, f-AnoGAN, DAE, DDPM and pDDPM are presented. From left to right, input, reconstruction, errormap, a heatmap of the errormap and the ground truth annotation is shown

Appendix B. Visualization of different noise levels

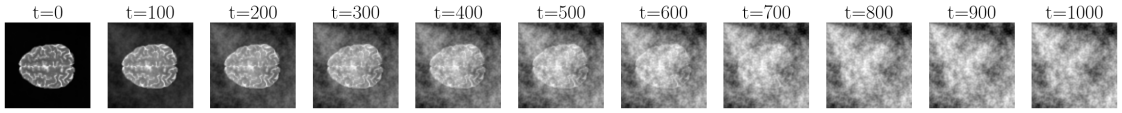


Figure 5: Training image from the IXI data set pertubed by simplex noise for different time steps $t = 0, 100, ..., 1000$

Appendix C. Ablation Studies for MSLUB

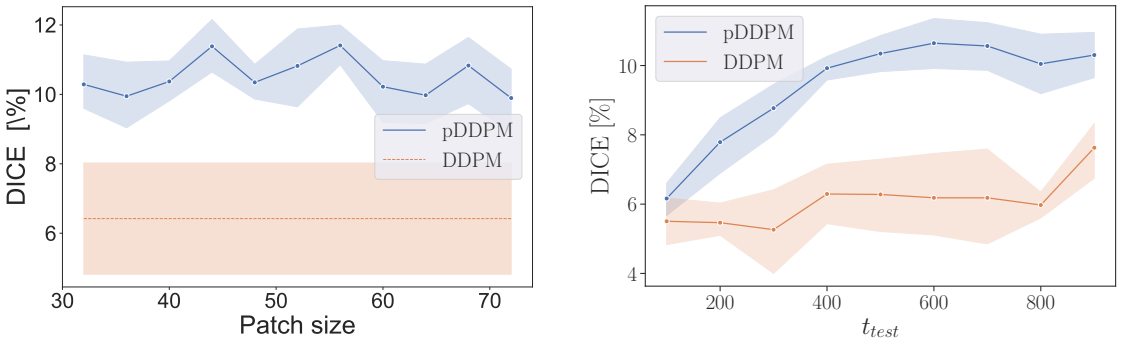


Figure 6: DICE for different patch sizes (left) and noise levels at test time t_{test} (right) for the MSLUB data set. We report the average DICE across the 5 cross-validation folds. Standard deviations are visualized as enveloping intervals.

8.3 Guided Reconstruction with Conditioned Diffusion Models for Unsupervised Anomaly Detection in Brain MRIs [22]

Computers in Biology and Medicine 186 (2025) 109660

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Guided reconstruction with conditioned diffusion models for unsupervised anomaly detection in brain MRIs

Finn Behrendt [a,*], Debayan Bhattacharya [a], Robin Mieling [a], Lennart Maack [a], Julia Krüger [b], Roland Opfer [b], Alexander Schlaefer [a]

[a] Hamburg University of Technology, Hamburg, Germany
[b] Jung Diagnostics GmbH, Hamburg, Germany

ARTICLE INFO

Keywords:
Unsupervised anomaly detection
Segmentation
Brain MRI
Diffusion models

ABSTRACT

The application of supervised models to clinical screening tasks is challenging due to the need for annotated data for each considered pathology. Unsupervised Anomaly Detection (UAD) is an alternative approach that aims to identify any anomaly as an outlier from a healthy training distribution. A prevalent strategy for UAD in brain MRI involves using generative models to learn the reconstruction of healthy brain anatomy for a given input image. As these models should fail to reconstruct unhealthy structures, the reconstruction errors indicate anomalies. However, a significant challenge is to balance the accurate reconstruction of healthy anatomy and the undesired replication of abnormal structures. While diffusion models have shown promising results with detailed and accurate reconstructions, they face challenges in preserving intensity characteristics, resulting in false positives. We propose conditioning the denoising process of diffusion models with additional information derived from a latent representation of the input image. We demonstrate that this conditioning allows for accurate and local adaptation to the general input intensity distribution while avoiding the replication of unhealthy structures. We compare the novel approach to different state-of-the-art methods and for different data sets. Our results show substantial improvements in the segmentation performance, with the Dice score improved by 11.9%, 20.0%, and 44.6%, for the BraTS, ATLAS and MSLUB data sets, respectively, while maintaining competitive performance on the WMH data set. Furthermore, our results indicate effective domain adaptation across different MRI acquisitions and simulated contrasts, an important attribute for general anomaly detection methods. The code for our work is available at https://github.com/FinnBehrendt/Conditioned-Diffusion-Models-UAD.

1. Introduction

Magnetic Resonance Imaging (MRI) is an important tool for diagnosing various conditions in the human brain [1,2]. However, interpreting brain MRIs can be error-prone, time-consuming, and places a significant workload on available radiologists [3,4]. To address these challenges and improve diagnostic efficiency, deep learning techniques like convolutional neural networks (CNN) have shown great promise in assisting radiologists by automating certain aspects of the analysis [2]. A common task is the detection and delineation of pathological structures in the MRI scans such as tumors [5], white matter lesions [6] or Alzheimer's disease [7]. Supervised deep learning approaches exhibit robust performance for these tasks, given that task-specific, annotated data sets are available. However, gathering such data sets is a cumbersome and costly process. Additionally, applying supervised models for screening tasks is difficult as any pathology must be detected, even those that are underrepresented or not included in the training data. In this context, screening tasks refer to scenarios where various conditions need to be identified without a specific target condition in mind. This includes assisting radiologists in detecting a wide range of findings, from expected abnormalities to unexpected or incidental ones, as well as applications in large population studies, such as the Hamburg City Health Study [8].

An alternative approach is unsupervised anomaly detection (UAD), which relies on healthy data instead of annotated pathologies. The goal is to learn the underlying data distribution of healthy brain MRI scans and to identify anomalies as outliers from that learned distribution. A popular approach is reconstruction-based UAD, where generative models (GM) are trained to reconstruct healthy anatomy. Given that the GMs are trained exclusively on healthy data, they should fail to generate unhealthy components of the input images and replace them

* Corresponding author.
E-mail address: finn.behrendt@tuhh.de (F. Behrendt).

https://doi.org/10.1016/j.compbiomed.2025.109660
Received 16 February 2024; Received in revised form 29 November 2024; Accepted 6 January 2025

by approximations of healthy anatomy. Subsequently, the discrepancy of input and pseudo-healthy reconstruction is measured, e.g., by the mean absolute error where large errors indicate anomalies [9–12]. Therefore, in theory, any deviation from the learned norm can be localized, including known conditions, unexpected anomalies such as artifacts, or previously unseen pathologies.

However, in practice, reconstruction-based UAD methods typically do not result in perfect anomaly detection and segmentation. Considering that the training task is reconstructing the input image, a key challenge is to limit the reconstruction to healthy brain anatomy. On the one hand, GMs that generate highly accurate reconstructions tend to perform a 'copy task', resulting in unhealthy structures still reflected in the pseudo-healthy reconstructions. These unhealthy structures do not deviate from the input image, resulting in false negatives in the segmentation mask. On the other hand, GMs with limited reconstruction accuracy may produce imperfect reconstructions everywhere, which in turn appear as differences in the anomaly map, even for healthy structures. This complicates distinguishing between actual pathologies and reconstruction errors, typically causing false positives in the segmentation map. Another related challenge in UAD is the potential domain shift between the distribution of the healthy data and the unhealthy test cases. Here, the reconstruction may appear different due to the domain shift; e.g., in simple cases, it is generally brighter than the input, and this shift reflects differences in the anomaly map. In summary, the challenge is to avoid copying unhealthy input features to the reconstruction while still adapting the general appearance of the reconstruction to the input.

Recently, denoising diffusion probabilistic models (DDPMs) [13] have shown promise as GMs for reconstruction-based UAD in brain MRI [14–16]. DDPMs generate images by denoising images corrupted by artificial noise, leveraging a high-dimensional latent space to preserve spatial context and achieve high-fidelity reconstructions. While the spatial latent space enables the accurate reconstruction of healthy structures, the additional denoising task aims to prevent the DDPMs from solely copying the content of the input image. Therefore, DDPMs can achieve a reasonable trade-off between the reconstruction accuracy of healthy and unhealthy structures [15,16]. However, a significant challenge remains in accurately reconstructing healthy brain anatomy that exhibits aligned intensity characteristics with the input image. The forward and backward processes of DDPMs do not adequately capture the highly variable local intensity distributions of MRI scans. This can result in discrepancies between the input and the reconstruction, which are difficult to distinguish from those arising from actual pathologies. This leads to reduced segmentation performance, particularly when facing domain shifts at test time. One potential solution could be to incorporate the input image as an additional input to the denoising process in the DDPM, e.g., as a second input channel. However, this could allow the denoising process to replicate the content of the input image, which would contradict the principles of reconstruction-based UAD.

In response to these challenges, we propose the use of conditioned DDPMs (cDDPMs) for UAD in brain MRI. Our approach includes a conditioning mechanism that guides the denoising process of the DDPM by utilizing an extra feature representation of the input image. This feature representation, derived from a CNN-based image encoder, does not capture detailed structural information. However, it contains the coarse local intensity information from the input image, which is often partially lost during the DDPM's forward process. This way, we aim to align the local intensity distribution of the reconstructed image without providing detailed structural information that could be used to replicate unhealthy structures. We conduct a comprehensive investigation of our conditioning approach, specifically addressing the challenges associated with reconstruction-based UAD in brain MRI. Initially, we evaluate the quality of healthy brain MRI reconstructions and assess the ability to reconstruct healthy anatomy while replacing unhealthy structures. Subsequently, we examine the domain adaptation capabilities of our method by assessing the alignment of intensity between input and reconstruction for datasets not encountered during training and by simulating varying contrast levels. Finally, we evaluate the segmentation performance of our approach on a variety of datasets, comparing established state-of-the-art UAD methods.

Our results indicate that our conditioning approach effectively aligns the local intensity distributions of input and reconstruction without supporting the replication of unhealthy structures. Furthermore, cDDPMs can effectively adapt to different intensity and contrast profiles of different MRI data sets. As a result, our proposed cDDPMs address key challenges of DDPMs in reconstruction-based UAD in brain MRI and improve or match the segmentation performance of the compared baseline models. When compared to DDPMs, the Dice score significantly increases ($p < 0.05$), rising from 50.27 to 56.30 for the BraTS21 dataset and from 20.18 to 24.22 for the ATLAS v2 dataset. For the MSLUB dataset, the performance improves from 9.71 to 14.04 and for the WMH dataset, both models report similar performance, with Dice scores of 12.06 and 11.59, respectively.

In summary, the main contributions of this work are:

- Reconstruction Quality: We develop a conditioning mechanism within cDDPMs that guides the denoising process of DDPMs while avoiding the replication of unhealthy structures.
- Domain Adaptation and Intensity Alignment: Our conditioning mechanism aligns the local intensity distribution of the reconstructed image with that of the input image. This effectively improves the generalization to intensity shifts of different MRI scans.
- Segmentation Performance: The accurate reconstructions and aligned local intensity distributions featured by our conditioning mechanism can significantly improve the segmentation performance and applicability of DDPMs for UAD in brain MRI.

This paper is organized as follows: In Section 2, we review relevant literature on UAD in brain MRI. In Section 3, we introduce DDPMs and subsequently explain our conditioning approach. In Section 4, we provide details of the experimental setup. In Section 5, we present the results and subsequently discuss them in Section 6. Finally, we provide a conclusion in Section 7.

2. Recent work

Autoencoders (AE) have been the primary focus of research on reconstruction-based UAD in brain MRI. Although these models exhibit potential in capturing the underlying healthy distribution, their effectiveness in UAD is limited by their blurry reconstructions [9]. To overcome this limitation, researchers have focused on improving the representations and reconstructions by adding skip connections with dropout [17], using multi-scale features [18], utilizing feature activation maps [19] or employing feature discrepancies [20,21]. Additionally, online outlier removal strategies [22] have been proposed for AEs. In parallel, Variational Autoencoders (VAE) have been investigated for the UAD task [23], focusing on enhancing the used context in 2D [24] and 3D [25,26] or utilizing restoration methods [11]. Moreover, Soft-Intro VAES (SIVAE) [27] have been investigated. Additionally, Generative Adversarial Networks (GAN) have been explored for the task of UAD either as pure GAN [28,29] or in combination with VAEs [30,31] and VQ-VAEs with transformers [12].

While AEs with skip connections and a spatial latent space enable reconstructions of high fidelity, they tend to perform a 'copy task' which enables the reconstruction of unhealthy anatomy and therefore contradicts the UAD principle [9,27]. Lately, [10] have shown that AEs with skip connections can be effectively used for UAD in brain MRI if they are regularized by an additional denoising task. Taking a similar direction, DDPMs have shown promise in the field of UAD [14,15,32,33]. DDPMs provide high reconstruction fidelity, but important information about the input image can be lost due to the noising

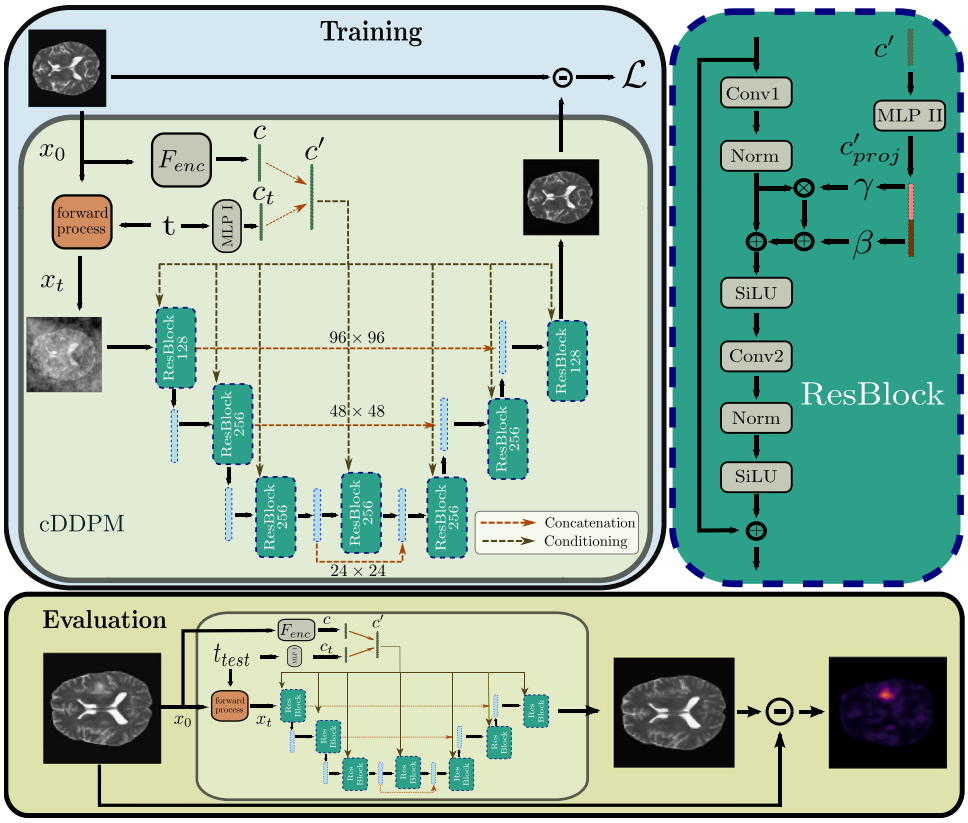


Fig. 1. Overview of our proposed approach. The encoder representations are learned along with the DDPM in the main training stage to condition the denoising process. The timestep embedding c_t is concatenated with the input image's projected encoder representation c. The resulting conditioning vector c' is used to scale and shift feature maps of the denoising Unet in the residual blocks. Each residual block consists of two convolution operators (conv1, conv2), group normalization (Norm) and Sigmoid Linear Units (SiLU). During evaluation, the residual map between unhealthy brain images and their healthy reconstructions is used for anomaly detection.

process. To address this, [16] proposed patch-based DDPMs that allow the use of parts of the original image content to provide information for the reconstruction of the input image. However, this patching strategy increases complexity and computational effort and can lead to artifacts in overlapping patch regions. A more efficient approach is seen in conditioning the denoising process of DDPMs with knowledge of the input image. Conditioned DDPMs have been successful in text-to-image synthesis tasks [34] and image-guided synthetic image generation [35, 36]. However, in the specific case of UAD, the objective is not to generate new images or to transfer styles but to accurately estimate a given input image while ensuring that unhealthy anatomy is absent in the estimation. Directly conditioning DDPMs with information from the input image can pose a risk of reconstructing unhealthy anatomy. Recent studies leverage different approaches to fuse information from a target image into the generation process of DDPMs. [37] introduce a classifier model trained to differentiate between healthy and tumorous brain MRI scans. At its core, this model employs classifier guidance during the sampling process, targeting the denoising towards a 'healthy' classification. Similarly, [38] apply classifier conditioning for counterfactual generation to remove tumors by conditioning the model to generate 'healthy' images. Despite showing promise, these methods rely on sample-level label information of given pathologies, contradicting the unsupervised setting. Moreover, their reliance on a scalar variable for conditioning limits the capacity to capture complex image characteristics. The authors [39] approach this challenge by incorporating synthetic anomalies in two separate diffusion processes. Their method involves feeding concatenated noised images (normal and anomalously altered) into a denoising Unet. This approach relies on synthetic defects that can be added to the input images. While effective in industrial defect detection, as demonstrated on the MVTEC dataset, this approach's dependency on synthetic anomalies raises concerns about its applicability to medical datasets. [40] propose an approach to fuse information of a target image into the generation process. By estimating and applying the same noise pattern to both the input and target images, their model aims to minimize the variance between these noised versions. While guiding the generation towards a target image, this method results in the loss of input image information, as the same noise is applied to the target image.

Our approach, detailed in the following sections, addresses these limitations by introducing a simple conditioning mechanism applied to the denoising Unet within DDPMs. In contrast to [37,38], we utilize spatial conditioning information and rely on unlabeled data. Unlike the proposed conditioning in [39], our approach does not rely on simulated anomalies and we directly condition the denoising process without the need for additional sub-networks. This direct conditioning of the denoising process in the Unet diverges from [40]'s approach, ensuring that valuable information from the target image is retained and utilized effectively throughout the diffusion process.

3. Methods

In our proposed cDDPM we use an image encoder network and embed the input image in a context vector $\boldsymbol{c} \in \mathbb{R}^d$ to condition the denoising Unet on meaningful features of the input image. Our motivation is that the additional information in $\boldsymbol{c}$ guides the generation process towards consistent intensity characteristics across the input image and its reconstruction. Hence, by introducing the context vector $\boldsymbol{c}$ we aim to recover local intensity information lost during the forward (noising) process of DDPMs. We utilize an image encoder with a dense latent space to extract information regarding the coarse shape and local intensity information of the noise-free input image. This latent representation can then be used to condition the denoising process and

supplement the individual context of the input image without providing detailed pixel-wise information that could be used to perform a 'copy task'. A general depiction of our approach is shown in Fig. 1.

3.1. DDPMs

DDPMs are generative models that learn the underlying data distribution of images $\boldsymbol{x} \in \mathbb{R}^{H,W,C}$ with height H, width W and C channels, given a training set. Training of DDPMs consists of two steps. The forward process, where an input image $\boldsymbol{x}_0$ is gradually transformed to Gaussian noise $\boldsymbol{x}_T = \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the backward process, where reversing the forward process is learned.

In the forward process, transforming $\boldsymbol{x}_0$ to $\boldsymbol{x}_T$ follows a predefined schedule $\beta_1, \dots, \beta_T$, where intermediate versions $\boldsymbol{x}_t$ are derived as

$$\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}),$$

$$\text{with } \bar{\alpha}_t = \prod_{s=0}^{t}(1-\beta_t).$$

The time step t controls the amount of added noise and is sampled from $t \sim Uniform(1, \dots, T)$. For edge cases, the image $\boldsymbol{x}_t$ is transformed to pure noise ($t = T$) or no transformation is applied ($t = 0$). In the backward process, the reconstructed image $\boldsymbol{x}_0^{rec}$ is recovered from $\boldsymbol{x}_t$ by

$$\boldsymbol{x}_0^{rec} \sim p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t),$$

$$\text{with } p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)).$$

Here, following [13], $\boldsymbol{\mu}_\theta$ is estimated by a Unet [41] with trainable parameters θ, and $\boldsymbol{\Sigma}_\theta(t) = \boldsymbol{\Sigma}(t) = \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t\mathbf{I}$ is fixed. Variational inference is used to achieve a tractable loss function and the variational lower bound (VLB) is derived as

$$\mathcal{L}_{VLB} = -log(p_\theta(\boldsymbol{x}_0))$$
$$+ D_{KL}(q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)).$$

which can be reformulated to

$$\mathcal{L}_{simple} = \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\|^2$$

by applying simplifications and by conditioning the denoising step on $\boldsymbol{x}_0$, as shown in [13]. In our work, instead of predicting the noise $\boldsymbol{\epsilon}$ we perform the equivalent task of directly estimating $\boldsymbol{x}_0^{rec} = \boldsymbol{x}_t - \boldsymbol{\epsilon}$. Hence, we derive our loss function as

$$\mathcal{L}_{rec} = |\boldsymbol{x}_0 - \boldsymbol{x}_0^{rec}|.$$

Typically, to generate new images with DDPMs, the backward step is applied step-wise to gradually denoise a random noise vector. For the given UAD task, we do not aim to generate new images but to estimate healthy brain anatomy given an input image. Therefore, we directly estimate $\boldsymbol{x}_0^{rec}$ given $\boldsymbol{x}_t$ at test time as it is done in [16]. The time step $t_{test} < T$ controls the level of noise to remove from $\boldsymbol{x}_t$ at test time. Optionally, to become agnostic to the noise magnitude, we use an ensemble of different values $t_{test} = [250, 500, 750]$ and average the reconstructions of each noise level, similar to [32].

3.2. Conditioned DDPMs (cDDPMs)

A general depiction of our conditioning approach is provided in Fig. 1. Formally, we condition the backward process of DDPMs on a context vector $\boldsymbol{c}$ as follows

$$\boldsymbol{x}_0^{rec} \sim p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{c}),$$

$$\text{with } p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{c}) = \mathcal{N}(\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}), \boldsymbol{\Sigma}(t)).$$

We use an image encoder F_{enc} to achieve a latent representation $\boldsymbol{c} = F_{enc}(\boldsymbol{x}_0)$ of the input image $\boldsymbol{x}_0$ where $\boldsymbol{c} \in \mathbb{R}^d$ with d as conditioning dimension.

To integrate the context vector $\boldsymbol{c}$, we manipulate the denoising Unet of the DDPM. Therefore, we individually adapt the features $\boldsymbol{f}_i \in \mathbb{R}^{H_i,W_i,C_i}$ at each level of the denoising Unet based on $\boldsymbol{c}$ where H_i, W_i and C_i are the respective feature map dimensions. To achieve this, we adapt the time step conditioning of DDPMs as follows. First, the time step is encoded using a sinusoidal position embedding. Next, t is projected to a vector $\boldsymbol{c}_t \in \mathbb{R}^d$ by a multi-layer perceptron (MLP I). Subsequently, we concatenate the context vector $\boldsymbol{c}$ and the time step vector $\boldsymbol{c}_t$, resulting in a conditioning vector $\boldsymbol{c}' \in \mathbb{R}^{2\cdot d}$. Finally, $\boldsymbol{c}'$ is projected to $\boldsymbol{c}'_{proj} \in \mathbb{R}^{2\cdot C_i}$ by another multi-layer perceptron (MLP II) at each feature level i. The vector $\boldsymbol{c}'_{proj}$ is then split into half, where the first and last C_i elements resemble the scaling factor γ and the shift value β. Inspired by [42], the variables γ and β are used to transform the individual feature maps as $\boldsymbol{f}'_i = \boldsymbol{f}_i * (\gamma + 1) + \beta$ in each residual block. This transformation adaptively scales and shifts the feature maps at each level of the denoising UNet, based on the context vector $\boldsymbol{c} = F_{enc}(\boldsymbol{x}_0)$, which encodes relevant information from the input image. The purpose of this transformation is to allow the model to dynamically adjust feature representations in response to both the image features and the conditioning information. By modulating the feature maps through context-based scaling and shifting, the model can more effectively preserve critical details of the clean target image while denoising the noisy input.

Optionally, to achieve a meaningful starting point for the calculation of the context vector $\boldsymbol{c} = F_{enc}(\boldsymbol{x}_0)$, we pre-train the feature extraction of the image encoder F_{enc} which is described in the next section.

3.3. Pre-training

We utilize a generative pre-training strategy for F_{enc}. More precisely, we utilize masked pre-training where typically transformer-based AEs are trained to reconstruct an image where a significant fraction of patches are masked out [43]. We adopt the SparK framework [44], where sparse convolutions and hierarchical features are used to enable the masked pre-training for CNNs. We pre-train the encoder with the same healthy training set as the main training task to learn the general feature representations required to capture important information from the MRI scans. After the pre-training stage, we discard the decoder and only use F_{enc} and fine-tune it along with the denoising Unet during the main training stage of the cDDPM. A schematic description of the pre-training stage is provided in Fig. 2.

4. Experimental setup

4.1. Data sets

Following the principle of UAD, we train our models for the reconstruction task on healthy data only (IXI). At test time, we evaluate the models' anomaly detection ability on unhealthy test sets of various pathologies (BraTS21, ATLAS, MSLUB, WMH). We provide an overview of all available MRI scanner details for the different data sets in Table 5.

4.1.1. Training data

We use the publicly available IXI data set[1] as our healthy reference data set for training. This data set includes 560 3D brain MRI scans collected from three medical facilities. Of the training data, 158 samples are set aside for testing, while the remaining data is divided into 5 folds, each containing 358 training samples and 44 validation samples for cross-validation.

Note that only the healthy training and validation data is used during pre-training and training stage.

[1] https://brain-development.org/ixi-dataset/

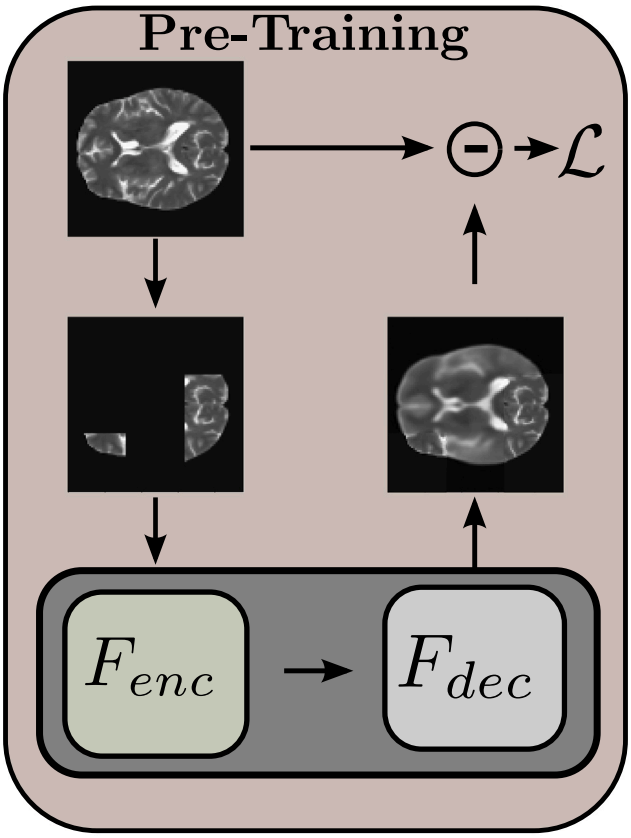


Fig. 2. Overview of the pre-training strategy. Random patches are erased from the input image. F_{enc} is used to derive a latent representation and F_{dec} is used to reconstruct the removed patches. After pre-training, the encoder F_{enc} is then fine-tuned along with the *DDPM* in the main training stage to condition the denoising process.

4.1.2. Evaluation data

For evaluation, we utilize four different publicly available data sets that contain different types of pathologies and the corresponding manual expert annotations:

1. Multimodal Brain Tumor Segmentation Challenge 2021 (BraTS21) [45–47]
2. Multiple Sclerosis data set from the University Hospital of Ljubljana (MSLUB) [48]
3. Anatomical Tracings of Lesions After Stroke v2.0 (ATLAS) [49]
4. White Matter Hyperintensity (WMH) [50]

The BraTS21 data set includes 2040 3D brain routine MRI scans of patients with glioma with a pathologically confirmed diagnosis. Accompanying the MRI scans, annotations from expert neuroradiologists are provided for 1251 scans that delineate tumor sub-regions as categorical masks. We fuse all sub-regions to obtain a binary segmentation mask to evaluate the anomaly detection task. All scans are available as T1-weighted volumes with and without contrast enhancement (T1-CE, T1) and T2-weighted or T2 fluid-attenuated inversion recovery (T2, FLAIR) volumes. The MSLUB data set includes 3D brain MRI scans of 30 patients with multiple sclerosis (MS) lesions. For each patient, along with the T1, T2 and FLAIR MRI scans, ground truth annotations are available derived based on multi-rater consensus. The ATLAS data set consists of 655 T1-weighted MRI scans of stroke patients collected from 44 research cohorts. The stroke lesions are annotated by domain experts and binary segmentation masks are provided. The WMH data set consists of 60 MRI scans of patients with white matter hyperintensities from three different institutions and scanner types. WMH segmentation masks are derived from the consensus of two expert radiologists. The datasets contain images acquired with different MRI parameters. While BraTS21 and MSLUB include T1, T2, and FLAIR data, WMH contains only T1 and FLAIR data, and ATLAS only T1 data. The IXI data set used for training contains T1 and T2 data. Hence, we train the models separately on T1 and T2 data, and evaluate on the respective data sets.

4.2. Pre-processing

We pre-process the images according to established pre-processing strategies for UAD in brain MRI [9]. First, we resample all MRI scans to the isotropic resolution of $1\,\text{mm} \times 1\,\text{mm} \times 1\,\text{mm}$ using cubic spline interpolation. Second, we register all MRI scans to the SRI24-Atlas. Third, we remove the skull from the MRI scans by skull stripping with HD-BET [51]. Subsequently, we crop each brain scan using its corresponding brain mask, removing unnecessary background while preserving relevant brain tissue. We then apply N4 bias field correction [52]. To standardize image sizes, we identify the largest brain volume in our dataset, add a safety margin, and pad all images to a unified size of $192 \times 192 \times 160$. Finally, to save computational resources, we reduce volume resolution by a factor of two and remove the 15 top and bottom slices parallel to the transverse plane, leading to a final resolution of $96 \times 96 \times 50$ voxels.

4.3. Post-processing

At test time, we derive a binary segmentation map from the residual map $\boldsymbol{R} = |\boldsymbol{x}_0 - \boldsymbol{x}_0^{rec}|$ where regions of high residuals indicate anomalies. To binarize $\boldsymbol{R}$, we first apply the following post-processing steps that are commonly used in the field of UAD in brain MRI [9,10,16,24]. First, a median filter with a kernel size of $K = 5 \times 5 \times 5$ is applied to smooth the residual map and to remove false positives. Subsequently, we perform brain mask eroding for 3 iterations. This step is mainly applied to filter out residuals due to poor reconstructions at sharp edges near the brain mask [9]. After binarization, we use connected component filtering to remove areas that include less than 7 voxels. This post-processing step removes false-positive predictions smaller than the anomalies considered in this study. In the supplemental material, we systematically compare the post-processing strategies for each data set.

4.4. Implementation details

This study compares our proposed cDDPM method with multiple established baselines for UAD in brain MRI. The baselines include *AE*, *VAE* [9], *SVAE* [26], Reverse Anomaly *RA* [27], *PHANES* [31] and denoising AEs *DAE* [10]. We also compare with simple thresholding *Thresh* [53]. Furthermore, we compare the feature-based reverse distillation method *RD* [21], Feature-Autoencoders *FAE* [20], the Encoder–Decoder Contrast method *EDC* [54] and the self-supervised Poisson Image Interpolation *PII* [55]. For a direct comparison, we also include *DDPM* [15] and patched DDPMs *pDDPM* [16] as a counterpart to our proposed method.

We adapt the baseline implementations by tuning hyper-parameters based on the validation set to improve training stability and performance. We set β_{VAE} to 0.001 for *VAE* and *SVAE*. For *EDC*, we use a lr of 1e−5 and 5e−5 for the encoder and decoder, respectively. *RA* and *PHANES* are trained with $\beta_{rec} = 16$ for the Soft-Intro VAE. For *PHANES*, we use a masking threshold of 0.15 and 0.10 for training and testing, respectively.

For *DDPM*, *pDDPM* and cDDPM, we use structured simplex noise, which has been shown to improve the UAD performance on MRI images [15]. Furthermore, we uniformly sample $t \in [1, T]$ with $T = 1000$ during training. At test time, we either use a fixed value of $t_{test} = \frac{T}{2} = 500$ or an ensemble of different values $t_{test} = [250, 500, 750]$ and average the individual reconstructions of each noise level. The denoising network for all DDPM-based approaches is an Unet similar to [34], with channel dimensions of [128, 256, 256]. As encoder network F_{enc},

we utilize a ResNet-backbone with a fully connected layer to match the target dimension of $c \in \mathbb{R}^d$ with $d = 128$ as conditioning dimension. We evaluated several conditioning dimensions, $d \in \{32, 64, 128, 256, 512\}$, and observed that the model achieved optimal performance when $d = 128$. A mask-out ratio of 65% is used during pre-training. For data augmentation, we utilize random -blur (p = 0.25), -bias (p = 0.25), -gamma (p = 0.5) and -ghosting (p = 0.5) from the torchio library [56]. We train for a maximum of 1600 epochs on NVIDIA V100 (32 GB) GPUs, using Adam as an optimizer, a learning rate of $1e{-}4$, and a batch size of 32. The best model checkpoint, as determined by performance on the validation set, is used for testing. The volumes are processed slice-wise, uniformly sampling slices with replacement during training and iterating over all slices to reconstruct the full volume at test time. We implement all models in Pytorch (v0.10).

4.5. Experiments and evaluation

We conduct various experiments to assess the individual features of our proposed cDDPMs. First, we evaluate the reconstruction quality for both healthy and unhealthy structures. Second, we investigate the generalization to real and simulated domain shifts between the training and test data. Lastly, we assess the final segmentation performance. For all experiments, we compare our cDDPMs to established baseline models on different data sets. The following subsections provide the detailed experimental settings for each individual evaluation.

4.5.1. Reconstruction quality

To evaluate the overall reconstruction quality, we utilize the held-out test set of the healthy IXI data set and calculate similarity metrics between input and reconstruction. We consider the Structural Similarity Index Measure (SSIM) [57], the Peak Signal To Noise Ratio (PSNR) and the Learned Perceptual Image Patch Similarity (LPIPS) as metrics to assess the reconstruction quality. For the feature-based LPIPs metric [58], features are extracted by a ResNet-based network, pre-trained on 3D medical data [59]. Furthermore, we report the reconstruction error ($l1$-error) for the healthy data set. For UAD, only healthy anatomy should be reconstructed. Hence, it is interesting to consider the $l1$-error of healthy and unhealthy anatomy separately, given the unhealthy evaluation data sets. Therefore, we calculate the $l1$-error for both healthy and unhealthy anatomy, as indicated by the annotation masks and calculate an $l1$-ratio as follows:

$$l1\text{-ratio} = \frac{l1_{unhealthy}}{l1_{healthy}}.$$

A higher value for the $l1$-ratio indicates that the model successfully reconstructs the healthy anatomy while struggling to reconstruct the unhealthy parts of the input, and vice versa.

4.5.2. Domain adaptation and intensity alignment

We investigate our proposed approach's generalization to domain shifts and the capability of adequately reconstructing the local intensity patterns of a given input image. We utilize data sets from different domains. For training, we utilize in-domain data from the IXI data set. As out-of-domain data, we utilize the BraTS21 data set. Notably, we only consider the content of the BraTS21 data set that is marked as healthy by the provided annotation masks, thereby ensuring the evaluation of domain shifts regarding scanner and brain diversity, not domain shifts arising from unhealthy brain MRI structures. Furthermore, we simulate different contrast levels ranging from $cl \in [0.3, 0.7, 1.0, 1.5, 2.0]$. The images of different contrast levels are derived by potentiating the gray values by the respective contrast level, i.e., $x_0^{cl=2} = x_0^2$. To evaluate the effect of the conditioning mechanism, we utilize the IXI data set and simulate different levels of conditioning information to investigate the reconstructions qualitatively. Therefore, we alter the available information in the image fed to the image encoder to condition the cDDPM. To achieve this, we crop the conditioning image at a given width of 50% and 100% where 100% indicates that the entire input image is used as the conditioning image. In addition to the qualitative evaluation of reconstructed, simulated data, we quantitatively assess the domain shift across input and reconstruction by comparing their intensity histograms. Therefore, we first calculate and plot the histograms. We partition the intensity values into 500 bins and divide the raw count by the total number of counts and the bin width. For a quantitative analysis of the distance between the intensity distributions, we calculate the Kullback–Leibler Divergence (KLD) as follows:

$$\text{KLD} = \left[-\sum_i p_{input} \log(p_{input}) \right] - \left[-\sum_i p_{reconstruction} \log(p_{reconstruction}) \right]$$

where $p = [p_1, p_2, \ldots, p_n]$ represents each intensity distribution.

4.5.3. Segmentation performance

To assess the segmentation performance for the UAD task, we utilize all unhealthy test sets and consider two different segmentation metrics. First, we report the best possible Dice score ($\lceil \text{DICE} \rceil$). The formula for the Dice score is given by:

$$\text{DICE} = \frac{2 \cdot |A \cap B|}{|A| + |B|}$$

where A and B are the predicted anomaly map and the ground truth annotation, respectively. The best possible $\lceil \text{DICE} \rceil$ is estimated by a greedy search of the threshold that optimizes the Dice score, similar to [60].

Second, we calculate the Area Under Precision–Recall Curve (AUPRC) as follows:

$$\text{AUPRC} = \sum_r (R(r) - R(r-1)) \cdot P(r).$$

Here, $R(r)$ represents the recall at a given threshold or rank r, and $P(r)$ represents the precision at the corresponding recall $R(r)$. The sum is taken over all thresholds or ranks r at which the precision and recall are computed.

4.5.4. Statistical testing

To conduct significance tests, we utilize the MLXtend library's permutation test [61] with 10,000 rounds of permutations and a significance level of $\alpha = 5\%$. This test computes the mean difference of the considered scores of two models for each permutation, and the resulting p-value is computed by counting the number of times the mean differences were equal to or greater than the sample differences, divided by the total number of permutations.

5. Results

We first compare the overall reconstruction quality of healthy and unhealthy structures. Second, we evaluate the domain adaptation properties of our approach and investigate the effect of our conditioning mechanism. Lastly, we evaluate the segmentation performance for different data sets, comparing our approach to established state-of-the-art UAD methods.

5.1. Reconstruction quality

In Table 1 we compare baseline models regarding their ability to reconstruct the healthy anatomy, given the held-out test set of the IXI data set. Overall, *DAEs*, *pDDPMs* and cDDPMs show high performance regarding the image-based similarity metrics. In contrast, lower reconstruction quality is reported for the dense autoencoder-based baselines. Comparing the DDPM-based approaches, both *pDDPM* and cDDPM outperform the baseline *DDPM* in terms of reconstruction quality with statistical significance ($p < 0.05$). We also analyze the unhealthy-to-healthy error ratio ($l1$-ratio) based on the unhealthy test

Table 1
Comparison of the reconstruction quality of the different models with the best results highlighted in bold. The asterisk * denotes superior performance with statistical significance compared to all other models ($p < 0.05$). For all metrics, the mean ± standard deviation across the different folds are reported. The arrows ↑ and ↓ indicate that higher and lower values are favorable, respectively. The $l1$-ratio is derived by dividing the $l1$-error of unhealthy anatomy by the $l1$-error of healthy anatomy. DDPM-based models are evaluated by ensembling different values for $t_{test} = [250, 500, 750]$.

Model	IXI (T2)				BraTS21 (T2)	MSLUB (T2)	ATLAS (T1)	WMH (T1)
	SSIM ↑	PSNR ↑	LPIPS (e-3) ↓	$l1$−error (e-3) ↓	$l1$−ratio ↑	$l1$−ratio ↑	l−ratio ↑	$l1$−ratio ↑
VAE [9]	74.98 ± 0.54	23.38 ± 0.14	4.03 ± 0.50	32.32 ± 0.64	3.52 ± 0.08	2.92 ± 0.06	4.43 ± 0.03	2.36 ± 0.04
SVAE [26]	77.87 ± 0.15	23.94 ± 0.06	3.31 ± 0.24	29.08 ± 0.16	3.90 ± 0.05	3.13 ± 0.05	3.38 ± 0.11	2.07 ± 0.01
AE [9]	76.11 ± 0.27	23.41 ± 0.14	3.19 ± 0.54	31.67 ± 0.41	3.84 ± 0.17	3.26 ± 0.18	4.40 ± 0.07	2.36 ± 0.04
RA [27]	75.46 ± 0.35	23.92 ± 0.23	2.18 ± 0.41	34.36 ± 1.43	3.10 ± 0.16	2.56 ± 0.11	3.93 ± 0.25	2.42 ± 0.19
PHANES [31]	69.04 ± 1.23	21.39 ± 0.32	1.08 ± 0.09	38.70 ± 1.74	3.54 ± 0.13	2.73 ± 0.07	4.01 ± 0.07	2.20 ± 0.05
DAE [10]	**98.69 ± 0.15***	**36.69 ± 0.38***	0.14 ± 0.01	**8.14 ± 0.17***	7.17 ± 0.63	2.69 ± 0.15	4.51 ± 0.15	2.99 ± 0.14
DDPM [15]	93.96 ± 0.37	31.79 ± 0.26	0.49 ± 0.14	14.29 ± 0.32	6.16 ± 0.53	3.37 ± 0.24	5.00 ± 0.23	**3.16 ± 0.15**
pDDPM [16]	96.62 ± 0.25	34.58 ± 0.39	**0.09 ± 0.04***	9.70 ± 0.43	7.16 ± 0.15	4.34 ± 0.13	5.58 ± 0.28	3.00 ± 0.16
cDDPM (Ours)	96.80 ± 0.19	34.87 ± 0.23	0.11 ± 0.05	9.68 ± 0.16	**7.43 ± 0.17**	**4.49 ± 0.18**	**5.69 ± 0.27**	3.12 ± 0.08

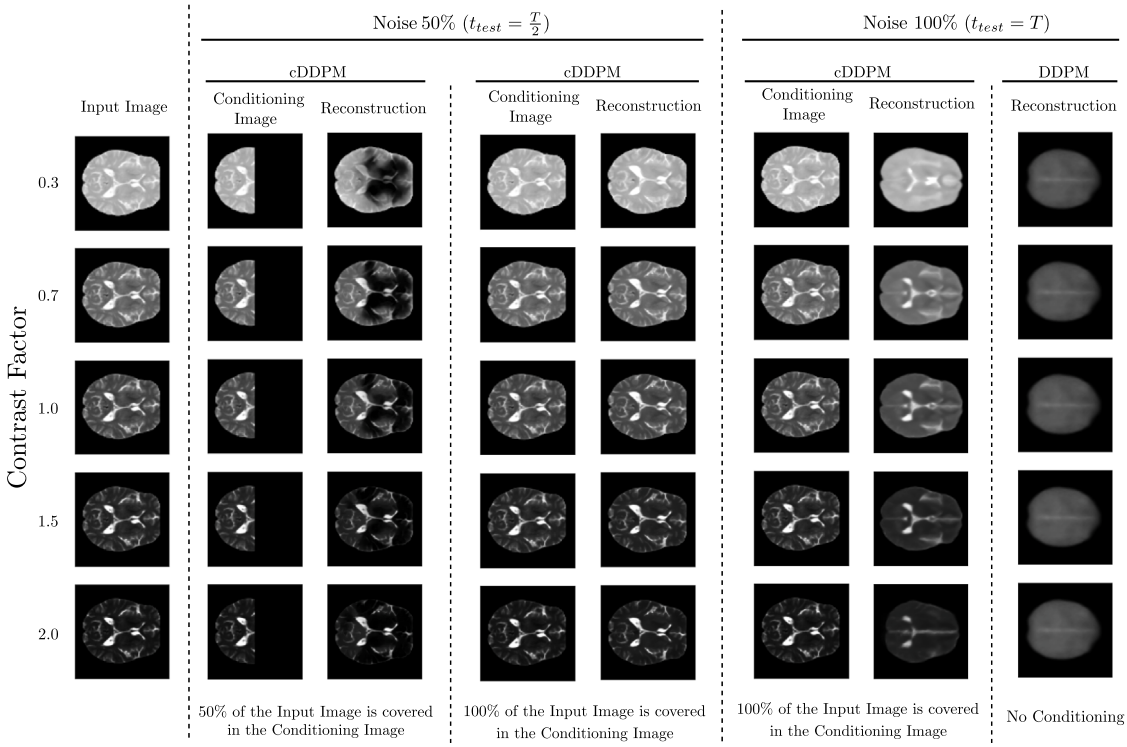


Fig. 3. Simulating the conditioning effect of the cDDPM for 50% of noise and 100% noise in the input image. In the first block, the input image that is fed to the *DDPM* or cDDPM is shown. In the second block, the reconstructions of cDDPM for different conditioning inputs are shown when a noise level of 50% is applied. In the third block, the reconstructions of cDDPMs and DDPMs are compared at a noise level of 100%. From top to bottom, the contrast level of the conditioning and input image is increased, respectively, for all columns.

sets. Generally, a higher $l1$-ratio is preferable as this indicates that the model successfully reconstructs the healthy anatomy without replicating the unhealthy parts of the input. Overall, the $l1$-ratio is highest for cDDPM across almost all data sets, except for the WMH, where *DDPM* shows competitive performance. While the $l1$-ratio of *DAEs* is high for the BraTS data set, it is substantially lower for all other data sets, indicating limited generalization. Additional results on the $l1$-error for all datasets are presented in Table 4 in the supplementary material.

5.2. Conditioning effect

To evaluate the effect the additional conditioning input has on individual reconstructions, we simulate different conditioning inputs, varying in the amount of used image information. Furthermore, we apply artificial contrast levels for the input images to mimic domain shifts. For each conditioning input and contrast level, we provide the reconstructions generated by cDDPMs in Fig. 3 for a noise level of $t_{test} = 500$ (50%). Moreover, we compare the reconstruction of DDPMs and cDDPMs in the extreme case of pure noise as input ($t_{test} = T = 1000$ (100%)). From Fig. 3, it becomes evident that the overall shape of the brain to reconstruct is preserved across the different (masked) conditioning images. However, the local intensity information in the reconstruction is only captured at regions covered in the conditioning image. In the extreme case of a noise level of 100%, the reconstructions of cDDPMs still coarsely follow the intensity distribution provided by the conditioning image, leading to a blurry reconstruction of the input image. In contrast, for unconditioned DDPMs, only a generic reconstruction that shares low similarity with the given input image can be obtained. Therefore, the conditioning mechanism primarily guides the generation process of cDDPMs to maintain local intensity distributions of the input image while detailed structural information is not captured.

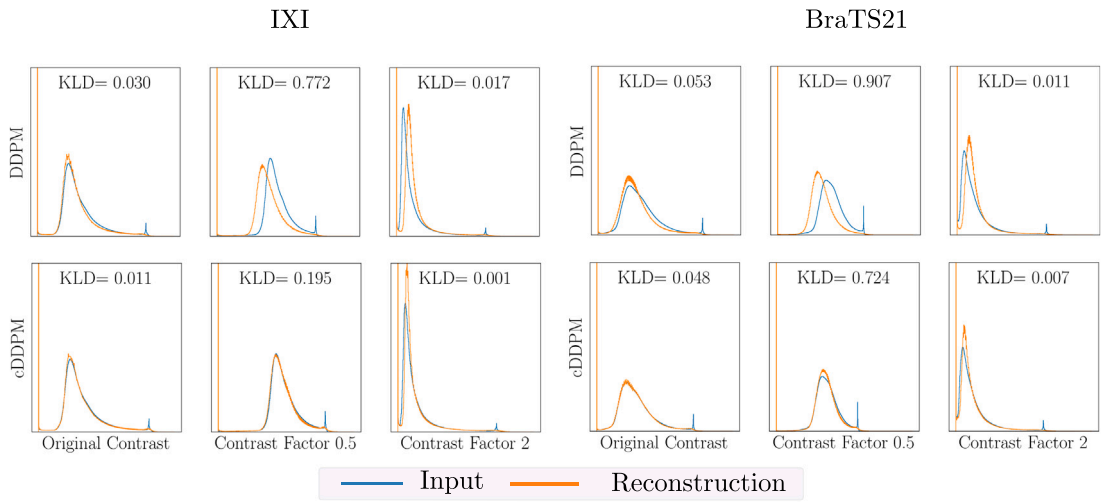


Fig. 4. Comparison of the histograms for input-reconstruction pairs of the healthy IXI (left) and the unhealthy BraTS21 (right) data set with original and augmented contrast. The top row shows the baseline *DDPM* without conditioning and the bottom row our proposed cDDPM with conditioning. The Kullback–Leibler divergence (KLD) for both histograms is indicated within each plot (lower values are preferable). Both models are evaluated by ensembling different values for $t_{test} = [250, 500, 750]$.

Table 2
Comparison of the evaluated models with the best results highlighted in bold. The asterisk * denotes superior performance with statistical significance compared to all other models ($p < 0.05$). For all metrics, the mean ± standard deviation across the different folds are reported. A checkmark at SSL denotes that a pre-trained encoder is used. A checkmark at ENS denotes the ensembling of different values for $t_{test} = [250, 500, 750]$. Otherwise, a fixed value of $t_{test} = 500$ is used for DDPM-based models.

Model	Modification		BraTS21 (T2)		MSLUB (T2)		ATLAS (T1)		WMH (T1)	
	ENS	SSL	[DICE] [%]	AUPRC [%]	[DICE] [%]	AUPRC [%]	[DICE] [%]	AUPRC [%]	[DICE] [%]	AUPRC [%]
Thresh [53]			30.26	20.27	7.65	4.23	4.66	1.71	10.32	4.72
VAE [9]			33.12 ± 1.12	25.74 ± 1.37	8.10 ± 0.18	4.48 ± 0.18	15.63 ± 0.73	11.44 ± 0.5	7.60 ± 0.28	3.86 ± 0.40
SVAE [26]			36.43 ± 0.36	30.3 ± 0.45	8.55 ± 0.11	4.8 ± 0.09	10.32 ± 0.53	6.84 ± 0.44	7.18 ± 0.07	2.97 ± 0.06
AE [9]			36.04 ± 1.73	28.8 ± 1.72	9.65 ± 0.97	5.71 ± 0.80	14.04 ± 0.6	10.16 ± 0.53	7.34 ± 0.08	3.43 ± 0.14
DAE [10]			48.82 ± 3.68	49.38 ± 4.18	7.57 ± 0.61	4.47 ± 0.69	15.95 ± 0.69	13.37 ± 0.62	12.02 ± 1.01	8.54 ± 1.02
RA [27]			16.75 ± 0.51	9.98 ± 0.43	3.96 ± 0.03	1.92 ± 0.04	12.21 ± 0.98	8.75 ± 0.93	6.04 ± 0.45	3.15 ± 0.31
PHANES [31]			28.42 ± 0.91	21.29 ± 1.06	6.11 ± 0.27	2.98 ± 0.07	17.62 ± 0.41	13.81 ± 0.48	7.55 ± 0.17	3.87 ± 0.13
RD [21]			32.57 ± 0.15	27.11 ± 0.15	6.48 ± 0.20	3.32 ± 0.06	19.69 ± 0.26	15.51 ± 0.20	7.48 ± 0.10	3.89 ± 0.07
FAE [20]			44.59 ± 2.19	43.63 ± 0.47	6.85 ± 0.65	3.85 ± 0.08	17.76 ± 0.16	13.91 ± 0.10	8.81 ± 0.38	4.77 ± 0.26
EDC [54]			36.66 ± 3.03	30.47 ± 4.25	7.23 ± 0.29	3.88 ± 0.4	18.67 ± 1.02	14.34 ± 0.86	8.62 ± 0.47	4.67 ± 0.37
PII [55]			40.83 ± 2.18	36.52 ± 2.66	9.46 ± 0.43	5.22 ± 0.37	9.73 ± 1.89	7.31 ± 1.64	6.59 ± 1.87	3.36 ± 1.03
DDPM [15]			49.43 ± 1.94	50.00 ± 2.13	9.63 ± 1.33	6.81 ± 1.54	17.57 ± 1.05	15.64 ± 0.90	11.56 ± 0.93	8.65 ± 0.87
DDPM [15]	✓		50.27 ± 2.67	50.61 ± 2.92	9.71 ± 1.29	6.27 ± 1.58	20.18 ± 0.58	17.77 ± 0.47	**12.06 ± 0.97**	8.89 ± 0.89
pDDPM [16]			53.25 ± 0.50	54.73 ± 0.52	12.40 ± 0.36	10.14 ± 0.44	19.20 ± 0.45	17.31 ± 0.34	10.14 ± 0.50	7.78 ± 0.56
pDDPM [16]	✓		53.61 ± 0.51	55.08 ± 0.54	12.83 ± 0.40	10.02 ± 0.36	19.92 ± 0.24	17.84 ± 0.10	10.13 ± 0.53	7.52 ± 0.56
cDDPM (Ours)			54.49 ± 1.63	56.81 ± 1.96	12.79 ± 1.08	10.07 ± 1.07	22.6 ± 1.67	20.65 ± 1.52	11.21 ± 0.54	9.05 ± 0.56
cDDPM (Ours)		✓	55.67 ± 1.05	58.14 ± 1.28	13.52 ± 0.91	10.89 ± 1.08	22.66 ± 1.20	20.85 ± 1.28	11.15 ± 0.8	9.03 ± 0.90
cDDPM (Ours)	✓	✓	**56.30 ± 1.25***	**58.82 ± 1.56***	**14.04 ± 1.16**	**10.97 ± 1.17**	**24.22 ± 1.10***	**22.22 ± 1.15***	11.59 ± 0.93	**9.26 ± 1.07**

5.3. Domain adaptation

To evaluate the domain adaptation to different data sets, we consider the healthy IXI data set as an in-domain data set and the unhealthy BraTS21 data set as an out-of-domain one. Note that for the BraTS21 data set, we only consider regions annotated as healthy. Thereby, we ensure the evaluation of domain shifts regarding scanner and brain diversity, not domain shifts arising from unhealthy brain MRI structures. In Fig. 4, we provide the histograms of input and reconstructions of the healthy IXI data set (left) and the unhealthy BraTS21 data set (right). We observe that DDPMs show substantial discrepancies across input and reconstruction intensity distributions. Particularly for simulated contrast levels, the histograms deviate. In contrast, the intensity distribution of images reconstructed by cDDPMs exhibits higher similarities with the input intensity distribution for in- and out-of-domain data. Compared to DDPMs, cDDPMs decrease the KLD by a factor of 2.3, 4.0 and 17.0 across the contrast factors, respectively, considering the IXI data set.

5.4. Segmentation performance

We report the segmentation performance in Table 2 considering all unhealthy test sets. For cDDPMs, improved performance is reported compared to all baselines across all data sets, except for the WMH data set, where the performance is on par with *DDPMs*. While the improvements for the cDDPM are statistically significant for the BraTS21 and ATLAS data sets ($p < 0.05$), no significant difference can be observed for the MSLUB and WMH data sets. Furthermore, we report enhanced performance of cDDPMs when pre-training the encoder (SSL checkmark in Table 2) and ensembling the reconstructions of different noise levels (ENS checkmark in Table 2) for most data sets. Notably, the inference time of cDDPMs is reduced by ∼ 37% compared to *pDDPMs* and increased by ∼ 2% compared to *DDPMs*.

The reconstruction-based *AEs*, *(S)VAEs*, *RA* and *PHANES* as well as the self-supervised *PII* and the feature-based *EDC* are outperformed by a margin when compared to DDPM-based approaches. While the feature-based methods *RD* and *FAE* and the reconstruction-based *DAEs* show competitive performance considering individual data sets, limited

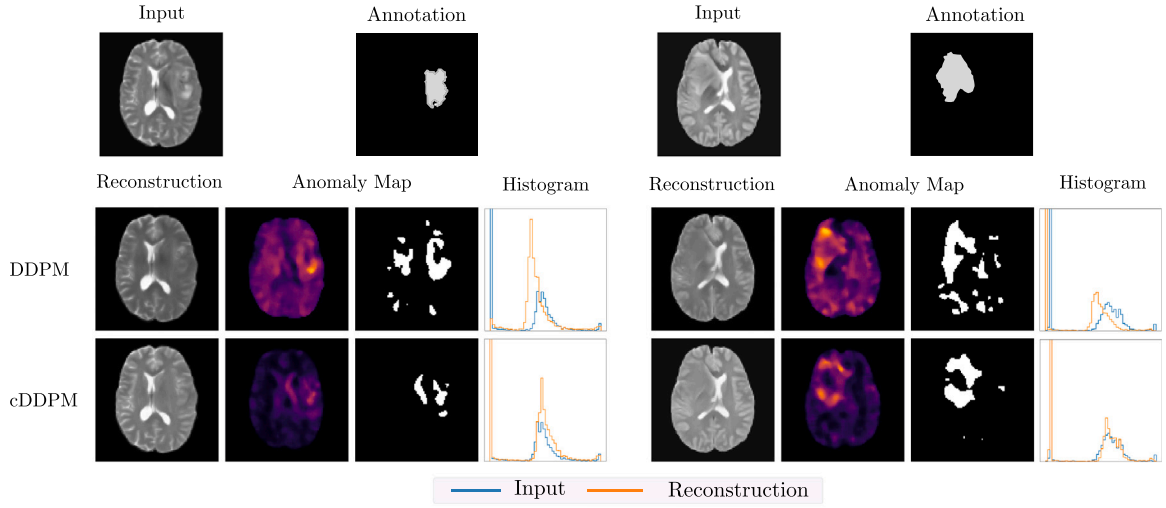


Fig. 5. Exemplary reconstructions and anomaly maps for DDPMs (second row) and cDDPMs (third row). The input and the corresponding ground truth annotation are provided in the first row. For each case, the reconstruction, the anomaly map and the histograms of intensity values in input and reconstruction are shown. Note that for histogram calculation, only healthy areas are considered. For visualization purposes, we provide segmentation maps next to the anomaly maps. We derive the binarization threshold by optimizing for the best possible dice score.

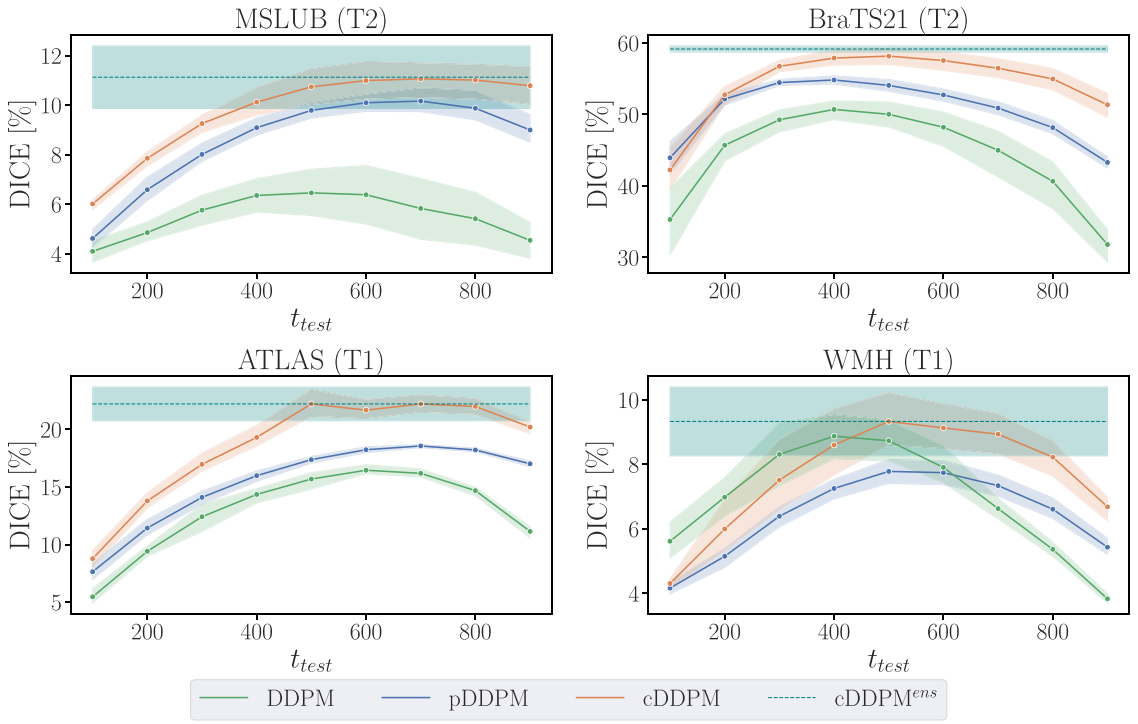


Fig. 6. Comparison of different noise levels t_{test} regarding the AUPRC. Top row: MSLUB (left) and BraTS21 (right) data sets. Bottom row: ATLAS (left) and WMH data set. The superscript *ens* denotes the ensembling of reconstructions from different noise levels t_{test} in $[250, 500, 750]$.

generalization across all evaluated data sets is shown. For a qualitative assessment of the conditioning mechanism, we provide reconstructions and the corresponding anomaly maps in Fig. 5, comparing unconditioned *DDPMs* to cDDPMs. cDDPMs provide accurate reconstructions of the target image with aligned intensity distributions across input and reconstructions. In contrast, the reconstruction of *DDPMs* shows less details with a slight intensity shift. Hence, the anomaly map of cDDPMs shows a higher contrast across normal and abnormal regions, which facilitates the delineation of the present pathology.

In Fig. 6 we provide an ablation study considering different noise levels applied during the diffusion process at test time. The achieved AUPRC score is dependent on the noise level. Applying the ensembling strategy reduces this dependency and achieves consistent performance without specifying an individual noise level.

We supply a collection of exemplary residual maps for different baseline methods in Fig. 7. Compared to the baselines, our proposed cDDPM demonstrates the low occurrences of false positives and a pronounced contrast in the residual maps. This observation is further supported by Figure 8 in the supplementary material.

Additionally, we include ablation studies on the applied post-processing steps in the supplementary material.

6. Discussion

UAD in brain MRI has gained significant attention due to its potential to identify abnormalities without costly data annotation. Compared to supervised approaches that rely on annotated data sets, UAD methods take a different approach by learning the underlying data distribution of healthy brain anatomy and identifying anomalies as outliers. This is a crucial property for screening tasks, where any pathology has to be detected, even if it is not represented in an annotated training set.

In this study, we focus on reconstruction-based UAD with DDPMs. DDPMs generate images by reconstructing an input corrupted by noise, leveraging the high-dimensional latent space to achieve high-fidelity reconstructions of fine-grained structures. However, we show that the forward and backward processes of DDPMs do not sufficiently capture the highly variable intensity characteristics of MRI scans. This results in differences between input and reconstructions that are difficult to distinguish from differences that arise from actual pathologies. This limitation becomes especially prominent in the presence of domain shifts at test time.

To address this challenge, we propose conditioned DDPMs (cDDPMs) for UAD in brain MRI. We train a DDPM to reconstruct healthy brain anatomy and condition the denoising process by a latent feature representation of the input image derived by an additional image encoder. While the additional dense feature representation can capture local intensity information of the image to reconstruct, it is unsuitable for the reconstruction of detailed structures [9,27]. This is important, as providing detailed structural information of the unhealthy input image could lead to copying abnormal structures, which would prevent the detection of pathologies. As demonstrated in our results, our approach facilitates the effective utilization of the conditioning signal, contributing to improved reconstructions that adapt locally to the input intensity distribution. Therefore, we observe enhanced domain adaptation capabilities to both real and simulated intensity profiles with our conditioning mechanism. Moreover, our results indicate that the additional conditioning signal does not support the replication of unhealthy structures. Finally, these individual features of our approach lead to an improved segmentation performance across various data sets.

In the following, we systematically evaluate our approach regarding reconstruction quality, domain adaptation, and segmentation performance based on five different data sets.

6.1. Reconstruction quality

We compare the reconstruction quality of our method with baseline models on the healthy IXI data set in Table 1. For the *AE* and *(S)VAE*, overall the worst reconstruction quality is reported. A reason for this is seen in the strict bottleneck enforced by the dense latent space as it inhibits information flow [9]. In contrast, methods like *DAE* or *DDPMs* that are not constrained by a dense latent space but by a denoising task [10] show improved reconstruction performance. While *pDDPMs* and cDDPMs outperform the baseline *DDPM* in terms of reconstruction quality, we observe that all models are outperformed by the *DAE*. We note that the overall training objective of the compared generative models is to reconstruct the image with high accuracy and copying the input image would be a trivial solution. However, for the UAD task, it is crucial that the given input image is not solely copied but that pseudo-healthy representatives replace unhealthy anatomy. Hence, comparing only the reconstruction quality of healthy anatomy does not necessarily reflect the usefulness of the UAD task. Therefore, we utilize the $l1$-ratio where high values indicate a better trade-off between the reconstruction of healthy and unhealthy anatomy and vice-versa. While *DDPMs* and particularly the cDDPM achieve a high $l1$-ratio, across all unhealthy data sets, it becomes evident that the *DAE* fails to generalize to different pathology types, a crucial property of UAD methods. A reason for that is seen in the chosen noise type in *DAE* that mimics the visual appearance of tumors [27,60]. In summary, the cDDPM shows improved reconstruction quality compared to *DDPMs* while preserving a high $l1$-ratio. We conclude that the conditioning mechanism effectively captures intensity information from the input image for an improved reconstruction without providing too much detailed structural information that would enable the cDDPM to solely copy the input image.

6.2. Domain adaptation

We evaluate the domain adaptation capabilities of cDDPMs by simulating different contrast levels and conditioning inputs in Fig. 3. The reconstructions show that while the overall shape is preserved across different conditioning masks, meaningful reconstructions are achieved only in regions covered by the conditioning image. Particularly, the conditioning image is critical in capturing local intensity information, demonstrating the ability of cDDPMs to adapt to different contrast levels and effectively capture intensity information. This becomes even more evident when high noise levels are considered (100%), where the only source of information concerning the given input image is the conditioning image. Here, the reconstruction becomes totally dependent on the shape and intensity characteristics of the conditioning image. The conditioning facilitates a blurred reconstruction of prominent local intensity changes. This indicates that the dense latent representation used for conditioning guides the reconstruction with local intensity details from the input image, while detailed structural information is not captured. Similar results are reported in the literature comparing AE architectures with dense and spatial latent spaces [9]. Furthermore, these findings indicate that cDDPMs effectively learn to balance information from the noisy input image and the conditioning encoder features during training, adapting according to the input's noise level.

We explore the domain adaptation capabilities of cDDPMs in real-world scenarios where a different, out-of-domain data set is used for testing. To assess the domain adaptation ability, we investigate the deviation between the intensity distributions of the input and reconstruction by plotting histograms and calculating the Kullback–Leibler Divergence (KLD) as a proxy in Fig. 4. Our results show that cDDPMs exhibit improved performance in capturing and estimating the intensity distribution. Particularly when simulating contrast levels considerably different from the training distribution, cDDPMs demonstrate superior alignment of the histograms and lower KLD values compared to unconditioned *DDPMs*. This analysis highlights the potential of the conditioning mechanism in cDDPMs to effectively adapt to unseen variations in intensity profiles and improve the coherence between input and reconstruction, which can contribute to improved domain adaptation.

6.3. Segmentation performance

Overall, cDDPMs demonstrate competitive or superior results compared to traditional autoencoder-based approaches, as well as the baseline *DDPM* and *pDDPM*, as presented in Table 2. Additionally, the feature-based and self-supervised methods are outperformed by our cDDPM. We conclude that the enhanced reconstruction quality, global and local intensity information capture, and effective domain adaptation capabilities attributed to our conditioning approach contribute to strong segmentation performance. Furthermore, we observe that pre-training the encoder F_{enc} slightly enhances the segmentation performance in most cases, indicating that starting from an already learned representation space has the potential to improve the overall integration of the conditioning features, compared to simultaneously training the parameters of both, DDPM and F_{enc} from scratch. Our results demonstrate improved performance of cDDPMs over *pDDPMs*, indicating that the proposed conditioning mechanism is more effective compared to the patching strategy in *pDDPMs*. A reason for this is seen in potential artifacts, introduced by the patching strategy. Furthermore, cDDPMs reduce complexity and inference time as there is no need for a

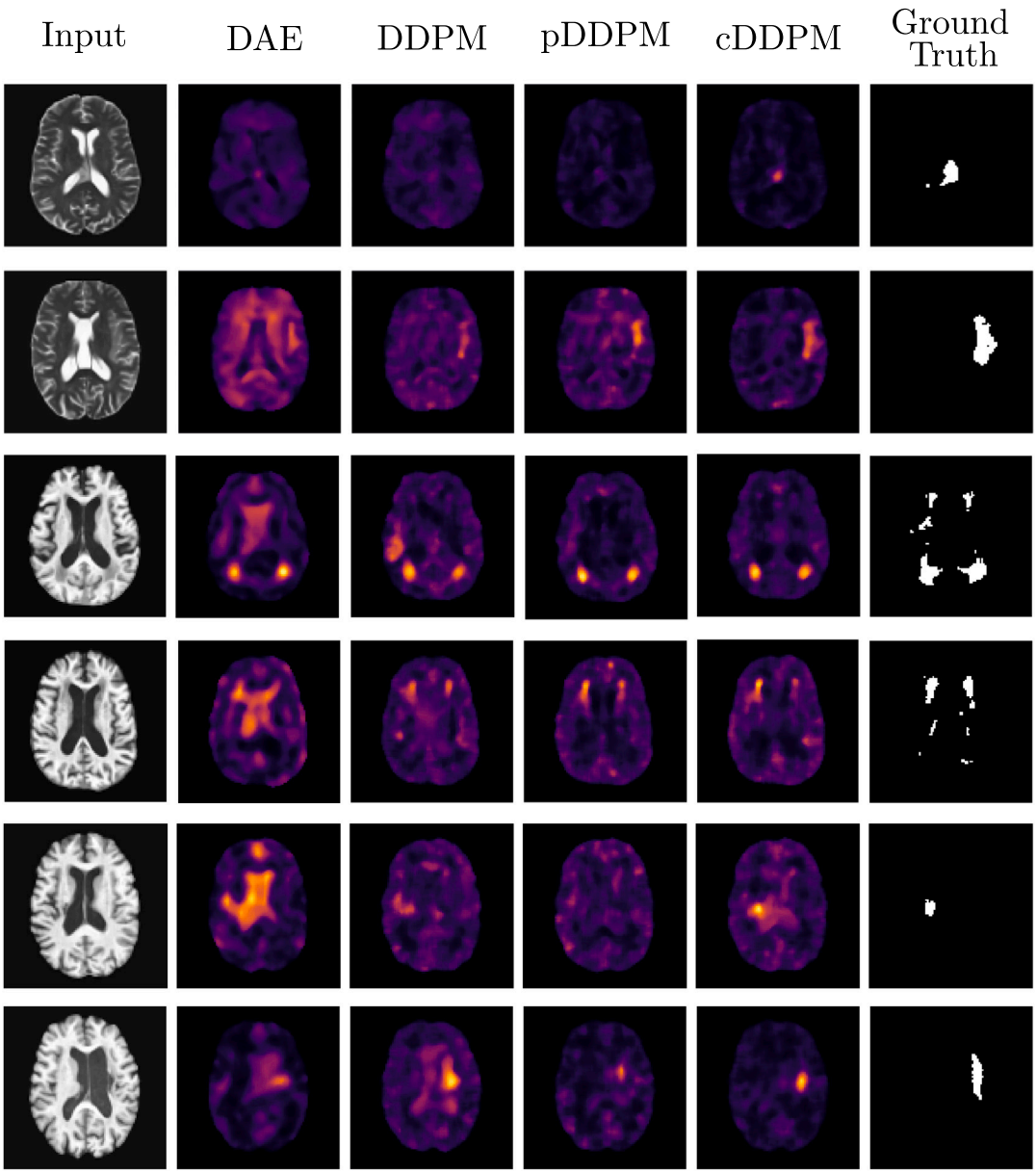


Fig. 7. Exemplary residual maps from the BraTS21 data set (rows 1 and 2), the WMH data set (rows 3 and 4) and the ATLAS data set (rows 5 and 6). The input image, residual maps, and ground truth are shown from left to right. Corresponding segmentation maps are provided in Figure 8.

costly patching strategy, making them a practical and efficient solution for UAD in brain MRI. Despite the superior reconstruction quality of the *DAE* on healthy data, cDDPMs show stronger segmentation results given the final UAD task. A reason for this is seen in the *DAEs*' ability to reconstruct unhealthy anatomy, particularly for pathologies differing from the appearance of tumors, as discussed in Section 6.1. Notably, the performance of *AEs* and *VAEs* in this work is lower compared to previous studies such as [9]. This discrepancy can be attributed to differences in the MRI sequences used. Unlike [9], which uses FLAIR images, we rely on T2- and T1-weighted images. Previous research has shown that hyperintense lesions in FLAIR images can often be detected using simple thresholding [53]. This indicates that the high performance of *AEs* and *VAEs* in [9] may be largely due to the hyperintensity of the lesions, rather than the model's reconstruction ability.

To further analyze the effect of the conditioning mechanism, we compare reconstructions and anomaly maps of *DDPMs* and cDDPMs in Fig. 5. In contrast to *DDPMs*, cDDPMs' reconstructions follow the local intensity information of the respective input images. This is in line with the observations in Fig. 4, where a lower distribution shift is reported for cDDPMs. This leads to a reduction of false positives, facilitating the delineation of anomalies.

In Fig. 6, we explore the impact of noise levels on the segmentation performance. We demonstrate that cDDPMs outperform the baseline models across different noise levels for most data sets. However, we also observe that the noise level is a crucial hyper-parameter. High noise levels tend to result in more blurry and generic reconstructions, whereas low noise levels enable sharper reconstructions, including unhealthy anatomy. Thus, selecting an appropriate noise level is essential to achieve reasonable performance. However, the optimal value for the applied noise depends on the evaluated data set, as shown in Fig. 6. We assume that the main reason for this dependency is the different sizes of pathologies in the considered data sets, as also stated in [33]. We apply different noise levels and average the resulting reconstructions to address this dependency. Thereby, we utilize complementary information

of different reconstructions and effectively mitigate the dependency on the noise level, which enhances the model's generalization abilities.

6.4. Limitations and future work

Despite showing promising results, our approach has limitations. Specifically, the segmentation performance falls below that of supervised algorithms. However, compared to supervised networks, UAD methods offer a crucial advantage in detecting any pathologies unseen during training. While additional efforts are required for the practical clinical application of UAD methods, our approach demonstrates superior performance to state-of-the-art UAD methods, adding a valuable contribution to the field of UAD in brain MRI.

Another avenue for future work is the incorporation of multi-scale image encodings into our conditioning mechanism. Our study does not utilize multi-scale analysis, which could be advantageous in capturing fine-grained details of the intensity patterns and contextual information at different resolutions. By carefully integrating multi-scale image encodings without allowing a copy of the conditioning image, we see the potential to enhance the performance of our cDDPMs in capturing both global and local features of the input images.

We conduct our studies using a downsampled resolution. To improve performance, especially for detecting smaller lesions, future work could explore and evaluate our proposed method at higher resolutions. A promising approach to balance the increased computational cost is the use of latent diffusion models [62]. To further enhance the conditioning mechanism, an additional direction for future work is to explore the use of 3D image encoders. In earlier studies, we have shown that 3D information can improve the reconstruction quality for VAEs [25,26] and expect similar improvements for DDPMs. Currently, our approach operates on 2D slices of the MRI data, which may limit the preservation of 3D context and spatial relationships between slices. By utilizing a 3D encoder to condition the 2D DDPM, we can potentially capture and preserve 3D contextual information.

7. Conclusion

UAD in brain MRI presents a promising alternative to supervised models, especially considering clinical screening tasks. DDPMs have demonstrated their utility for UAD, largely due to their high reconstruction accuracy of fine-grained structures. However, accurately reconstructing the intensity characteristics of a given MRI scan remains a challenge, especially when facing domain shifts. To address this, we propose cDDPMs and introduce a conditioning mechanism that incorporates an additional feature representation of the input image into the DDPMs' denoising process. Our findings indicate that this conditioning mechanism effectively addresses challenges of DDPMs regarding capturing accurate intensity capture and domain adaptation. As a result, our approach outperformed state-of-the-art architectures for UAD in brain MRI on various publicly available data sets. Our work addresses challenges of applying DDPMs for UAD in brain MRI and has practical implications for detecting and segmenting pathologies in scenarios where domain shifts are likely.

CRediT authorship contribution statement

Finn Behrendt: Writing – original draft, Software, Methodology, Conceptualization. **Debayan Bhattacharya:** Writing – review & editing, Methodology, Conceptualization. **Robin Mieling:** Writing – review & editing, Methodology, Conceptualization. **Lennart Maack:** Writing – review & editing, Methodology, Conceptualization. **Julia Krüger:** Writing – review & editing, Funding acquisition, Conceptualization. **Roland Opfer:** Writing – review & editing, Funding acquisition, Conceptualization. **Alexander Schlaefer:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Ethics statement

This study was conducted using publicly available, fully anonymized datasets, and no new data were collected or experiments involving human subjects were performed. All procedures adhered to relevant laws, institutional guidelines, and ethical standards for research.

Declaration of competing interest

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

Acknowledgments

This work was partially funded by the "Zentrales Innovationsprogramm Mittelstand, Germany" [grant numbers KK5208101KS0 and ZF4026303TS9] and by the Free and Hanseatic City of Hamburg (Interdisciplinary Graduate School), Germany.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2025.109660.

References

[1] M.W. Vernooij, M.A. Ikram, H.L. Tanghe, A.J. Vincent, A. Hofman, G.P. Krestin, W.J. Niessen, M.M. Breteler, A. van der Lugt, Incidental findings on brain MRI in the general population, N. Engl. J. Med. 357 (18) (2007) 1821–1828.

[2] A.S. Lundervold, A. Lundervold, An overview of deep learning in medical imaging focusing on MRI, Z. Med. Phys. 29 (2) (2019) 102–127, http://dx.doi.org/10.1016/j.zemedi.2018.11.002, URL https://www.sciencedirect.com/science/article/pii/S0939388918301181, Special Issue: Deep Learning in Medical Physics.

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

[4] R.J. McDonald, K.M. Schwartz, L.J. Eckel, F.E. Diehn, C.H. Hunt, B.J. Bartholmai, B.J. Erickson, D.F. Kallmes, The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload, Academic Radiol. 22 (9) (2015) 1191–1198.

[5] M. Perkuhn, P. Stavrinou, F. Thiele, G. Shakirin, M. Mohan, D. Garmpis, C. Kabbasch, J. Borggrefe, Clinical evaluation of a multiparametric deep learning model for glioblastoma segmentation using heterogeneous magnetic resonance imaging data from clinical routine, Invest. Radiol. 53 (11) (2018) 647.

[6] P. Moeskops, J. de Bresser, H.J. Kuijf, A.M. Mendrik, G.J. Biessels, J.P. Pluim, I. Išgum, Evaluation of a deep learning approach for the segmentation of brain tissues and white matter hyperintensities of presumed vascular origin in MRI, NeuroImage: Clin. 17 (2018) 251–262.

[7] J. Islam, Y. Zhang, Brain MRI analysis for alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks, Brain Informat. 5 (2018) 1–14.

[8] A. Jagodzinski, C. Johansen, U. Koch-Gromus, G. Aarabi, G. Adam, S. Anders, M. Augustin, R.B. der Kellen, T. Beikler, C.-A. Behrendt, et al., Rationale and design of the hamburg city health study, Eur. J. Epidemiol. 35 (2020) 169–181.

[9] C. Baur, S. Denner, B. Wiestler, N. Navab, S. Albarqouni, Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study, Med. Image Anal. (2021) 101952.

[10] A. Kascenas, N. Pugeault, A.Q. O'Neil, Denoising autoencoders for unsupervised anomaly detection in brain MRI, in: Medical Imaging with Deep Learning, in: Proceedings of Machine Learning Research, PMLR, 2022.

[11] X. Chen, S. You, K.C. Tezcan, E. Konukoglu, Unsupervised lesion detection via image restoration with a normative prior, Med. Image Anal. 64 (2020) 101713.

[12] W.H. Pinaya, P.-D. Tudosiu, R. Gray, G. Rees, P. Nachev, S. Ourselin, M.J. Cardoso, Unsupervised brain imaging 3D anomaly detection and segmentation with transformers, Med. Image Anal. 79 (2022) 102475.

[13] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, NIPS 33 (2020) 6840–6851.

[14] W.H. Pinaya, M.S. Graham, R. Gray, P.F. Da Costa, P.-D. Tudosiu, P. Wright, Y.H. Mah, A.D. MacKinnon, J.T. Teo, R. Jager, et al., Fast unsupervised brain anomaly detection and segmentation with diffusion models, 2022, arXiv preprint arXiv:2206.03461.

[15] J. Wyatt, A. Leach, S.M. Schmon, C.G. Willcocks, AnoDDPM: Anomaly detection with denoising diffusion probabilistic models using simplex noise, in: CVPR, 2022, pp. 650–656.

[16] F. Behrendt, D. Bhattacharya, J. Krüger, R. Opfer, A. Schlaefer, Patched diffusion models for unsupervised anomaly detection in brain mri, in: Medical Imaging with Deep Learning, PMLR, 2024, pp. 1019–1032.

[17] C. Baur, B. Wiestler, S. Albarqouni, N. Navab, Bayesian skip-autoencoders for unsupervised hyperintense anomaly detection in high resolution brain MRI, in: IEEE ISBI, IEEE, 2020, pp. 1905–1909.

[18] C. Baur, B. Wiestler, S. Albarqouni, N. Navab, Scale-space autoencoders for unsupervised anomaly segmentation in brain mri, in: Computer Assisted Radiology and Surgery, Springer, 2020, pp. 552–561.

[19] J. Silva-Rodríguez, V. Naranjo, J. Dolz, Constrained unsupervised anomaly segmentation, Med. Image Anal. 80 (2022) 102526.

[20] F. Meissen, J. Paetzold, G. Kaissis, D. Rueckert, Unsupervised anomaly localization with structural feature-autoencoders, 2022, arXiv preprint arXiv:2208.10992.

[21] H. Deng, X. Li, Anomaly detection via reverse distillation from one-class embedding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 9737–9746.

[22] F. Behrendt, M. Bengs, F. Rogge, J. Krüger, R. Opfer, A. Schlaefer, Unsupervised anomaly detection in 3D brain MRI using deep learning with impured training data, in: 2022 IEEE 19th International Symposium on Biomedical Imaging, ISBI, IEEE, 2022, pp. 1–4.

[23] D. Zimmerer, F. Isensee, J. Petersen, S. Kohl, K. Maier-Hein, Unsupervised anomaly localization using variational auto-encoders, in: D. Shen, T. Liu, T.M. Peters, L.H. Staib, C. Essert, S. Zhou, P.-T. Yap, A. Khan (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing, Cham, 2019, pp. 289–297.

[24] D. Zimmerer, S. Kohl, J. Petersen, F. Isensee, K. Maier-Hein, Context-encoding variational autoencoder for unsupervised anomaly detection, in: Medical Imaging with Deep Learning, 2019.

[25] M. Bengs, F. Behrendt, J. Krüger, R. Opfer, A. Schlaefer, Three-dimensional deep learning with spatial erasing for unsupervised anomaly segmentation in brain MRI, Comput. Assist. Radiol. Surg. 16 (9) (2021) 1413–1423.

[26] F. Behrendt, M. Bengs, D. Bhattacharya, J. Krüger, R. Opfer, A. Schlaefer, Capturing inter-slice dependencies of 3D brain MRI-scans for unsupervised anomaly detection, in: Medical Imaging with Deep Learning, 2022, URL https://openreview.net/forum?id=db8wDgKH4p4.

[27] C.I. Bercea, B. Wiestler, D. Rueckert, J.A. Schnabel, Generalizing unsupervised anomaly detection: Towards unbiased pathology screening, in: Medical Imaging with Deep Learning, 2023, URL https://openreview.net/forum?id=8ojx-Ld3yjR.

[28] C. Han, L. Rundo, K. Murao, T. Noguchi, Y. Shimahara, Z.Á. Milacski, S. Koshino, E. Sala, H. Nakayama, S. Satoh, MADGAN: Unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction, BMC Bioinformatics 22 (2) (2021) 1–20.

[29] T. Schlegl, P. Seeböck, S.M. Waldstein, G. Langs, U. Schmidt-Erfurth, f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks, Med. Image Anal. 54 (2019) 30–44.

[30] C. Baur, B. Wiestler, S. Albarqouni, N. Navab, Deep autoencoding models for unsupervised anomaly segmentation in brain MR images, in: MICCAI Brainlesion Workshop, Springer, 2018, pp. 161–169.

[31] C.I. Bercea, B. Wiestler, D. Rueckert, J.A. Schnabel, Reversing the abnormal: Pseudo-healthy generative networks for anomaly detection, in: Medical Image Computing and Computer Assisted Intervention – MICCAI23, vol. 14224, 2023, pp. 293–303, http://dx.doi.org/10.1007/978-3-031-43904-9_29.

[32] M.S. Graham, W.H. Pinaya, P.-D. Tudosiu, P. Nachev, S. Ourselin, M.J. Cardoso, Denoising diffusion models for out-of-distribution detection, 2022, arXiv preprint arXiv:2211.07740.

[33] C.I. Bercea, M. Neumayr, D. Rueckert, J.A. Schnabel, Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models, in: ICML 3rd Workshop on Interpretable Machine Learning in Healthcare, IMLH, 2023, URL https://openreview.net/forum?id=kTpafpXrqa.

[34] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, NIPS 34 (2021) 8780–8794.

[35] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, M. Norouzi, Palette: Image-to-image diffusion models, in: ACM, 2022, pp. 1–10.

[36] T. Wang, T. Zhang, B. Zhang, H. Ouyang, D. Chen, Q. Chen, F. Wen, Pretraining is all you need for image-to-image translation, 2022, arXiv preprint arXiv:2205.12952.

[37] J. Wolleb, F. Bieder, R. Sandkühler, P.C. Cattin, Diffusion models for medical anomaly detection, 2022, arXiv preprint arXiv:2203.04306.

[38] P. Sanchez, A. Kascenas, X. Liu, A.Q. O'Neil, S.A. Tsaftaris, What is healthy? generative counterfactual diffusion for lesion localization, in: MICCAI Workshop on Deep Generative Models, Springer, 2022, pp. 34–44.

[39] X. Zhang, N. Li, J. Li, T. Dai, Y. Jiang, S.-T. Xia, Unsupervised surface anomaly detection with diffusion probabilistic model, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 6782–6791.

[40] A. Mousakhan, T. Brox, J. Tayyub, Anomaly detection with conditioned denoising diffusion models, 2023, arXiv preprint arXiv:2305.15956.

[41] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer Assisted Intervention, Springer, 2015, pp. 234–241.

[42] E. Perez, F. Strub, H. De Vries, V. Dumoulin, A. Courville, Film: Visual reasoning with a general conditioning layer, in: AAAI, Vol. 32, No. 1, 2018.

[43] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: CVPR, 2022, pp. 16000–16009.

[44] K. Tian, Y. Jiang, Q. Diao, C. Lin, L. Wang, Z. Yuan, Designing BERT for convolutional networks: Sparse and hierarchical masked modeling, 2023, arXiv:2301.03580.

[45] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F.C. Kitamura, S. Pati, et al., The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification, 2021, arXiv preprint arXiv:2107.02314.

[46] B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imaging 34 (10) (2014) 1993–2024.

[47] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, J.B. Freymann, K. Farahani, C. Davatzikos, Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features, Sci. Data 4 (1) (2017) 1–13.

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

[49] S.-L. Liew, B.P. Lo, M.R. Donnelly, A. Zavaliangos-Petropulu, J.N. Jeong, G. Barisano, A. Hutton, J.P. Simon, J.M. Juliano, A. Suri, et al., A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms, Sci. Data 9 (1) (2022) 320.

[50] H.J. Kuijf, J.M. Biesbroek, J. De Bresser, R. Heinen, S. Andermatt, M. Bento, M. Berseth, M. Belyaev, M.J. Cardoso, A. Casamitjana, et al., Standardized assessment of automatic segmentation of white matter hyperintensities and results of the WMH segmentation challenge, IEEE Trans. Med. Imaging 38 (11) (2019) 2556–2568.

[51] F. Isensee, M. Schell, I. Pflueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H.-P. Schlemmer, S. Heiland, W. Wick, et al., Automated brain extraction of multisequence MRI using artificial neural networks, Hum. Brain Mapp. 40 (17) (2019) 4952–4964.

[52] N.J. Tustison, B.B. Avants, P.A. Cook, Y. Zheng, A. Egan, P.A. Yushkevich, J.C. Gee, N4ITK: improved N3 bias correction, IEEE Trans. Med. Imaging 29 (6) (2010) 1310–1320.

[53] F. Meissen, G. Kaissis, D. Rueckert, Challenging current semi-supervised anomaly segmentation methods for brain MRI, in: MICCAI Brainlesion Workshop, Springer, 2022, pp. 63–74.

[54] J. Guo, S. Lu, L. Jia, W. Zhang, H. Li, Encoder-decoder contrast for unsupervised anomaly detection in medical images, IEEE Trans. Med. Imaging 43 (3) (2023) http://dx.doi.org/10.1109/TMI.2023.3327720.

[55] J. Tan, B. Hou, T. Day, J. Simpson, D. Rueckert, B. Kainz, Detecting outliers with poisson image interpolation, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24, Springer, 2021, pp. 581–591.

[56] F. Pérez-García, R. Sparks, S. Ourselin, Torchio: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning, Comput. Methods Programs Biomed. 208 (2021) 106236.

[57] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612, http://dx.doi.org/10.1109/TIP.2003.819861.

[58] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.

[59] S. Chen, K. Ma, Y. Zheng, Med3d: Transfer learning for 3d medical image analysis, 2019, arXiv preprint arXiv:1904.00625.

[60] I. Lagogiannis, F. Meissen, G. Kaissis, D. Rueckert, Unsupervised pathology detection: A deep dive into the state of the art, IEEE Trans. Med. Imaging PP (2023) http://dx.doi.org/10.1109/TMI.2023.3298093.

[61] S. Raschka, Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack, J. Open Source Softw. 3 (24) (2018).

[62] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: CVPR, 2022, pp. 10684–10695.

Supplementary Material

Post-Processing Analysis

In Table 3, we provide an analysis of the applied post-processing steps by excluding individual post-processing steps from the evaluation protocol. We show that while the median filter shows to have a large effect, the other post-processing techniques only show minor changes. Moreover, no post-processing strategy consistently works for all models or data sets, motivating further research and a systematic study about the effect of different post-processing steps for UAD in brain MRI.

Table 3: Post-processing analysis for all data sets regarding AUPRC. ✓indicates the presence and ✗indicates the absense of Connected Component (CC), Medianfiltering (MF) or Brain Eroding (BE) in the evaluation phase, respectively. For all models, the mean ± standard deviation of the AUPRC are provided. Color-coded absolute differences concerning the respective baseline models are provided in the brackets.

Model	CC	MF	BE	BraTS21 (T2)	MSLUB (T2)	ATLAS (T1)	WMH (T1)
DAE	✓	✓	✓	49.38 ± 4.18	4.47 ± 0.69	13.37 ± 0.62	8.54 ± 1.02
DAE	✗	✓	✓	49.38 ± 4.18 (0.00)	4.47 ± 0.69 (0.00)	8.53 ± 0.24 (-4.84)	7.31 ± 0.91 (-1.23)
DAE	✓	✗	✓	39.69 ± 3.68 (-9.69)	3.92 ± 0.35 (-0.55)	9.21 ± 0.32 (-4.16)	7.01 ± 0.68 (-1.53)
DAE	✓	✓	✗	48.79 ± 4.31 (-0.59)	3.84 ± 0.55 (-0.63)	8.77 ± 0.28 (-4.60)	6.80 ± 0.81 (-1.74)
DDPM	✓	✓	✓	50.61 ± 2.92	6.27 ± 1.58	17.77 ± 0.47	8.89 ± 0.89
DDPM	✗	✓	✓	50.68 ± 2.81 (0.07)	6.25 ± 1.51 (-0.02)	14.65 ± 0.3 (-3.12)	10.27 ± 0.96 (1.38)
DDPM	✓	✗	✓	35.93 ± 2.27 (-14.68)	5.40 ± 0.95 (-0.87)	12.02 ± 0.45 (-5.75)	8.80 ± 0.71 (-0.09)
DDPM	✓	✓	✗	50.47 ± 3.07 (-0.14)	5.60 ± 1.31 (-0.67)	15.11 ± 0.32 (-2.66)	9.89 ± 1.00 (1.00)
pDDPM	✓	✓	✓	55.08 ± 0.54	10.02 ± 0.36	17.84 ± 0.10	7.52 ± 0.56
pDDPM	✗	✓	✓	55.07 ± 0.53 (-0.01)	10.06 ± 0.37 (0.04)	12.14 ± 0.21 (-5.70)	7.33 ± 0.39 (-0.19)
pDDPM	✓	✗	✓	34.99 ± 0.43 (-20.09)	7.35 ± 0.28 (-2.67)	13.35 ± 0.22 (-4.49)	7.62 ± 0.80 (0.10)
pDDPM	✓	✓	✗	54.10 ± 0.57 (-0.98)	8.63 ± 0.48 (-1.39)	13.71 ± 0.28 (-4.13)	7.3 ± 0.95 (-0.22)
cDDPM	✓	✓	✓	58.82 ± 1.56	10.97 ± 1.17	22.22 ± 1.15	9.26 ± 1.07
cDDPM	✗	✓	✓	58.84 ± 1.57 (0.02)	11.22 ± 1.31 (0.25)	19.92 ± 1.45 (-2.30)	9.86 ± 1.18 (0.60)
cDDPM	✓	✗	✓	39.70 ± 1.40 (-19.12)	8.31 ± 0.98 (-2.66)	16.56 ± 1.11 (-5.66)	8.34 ± 0.76 (-0.92)
cDDPM	✓	✓	✗	58.52 ± 1.62 (-0.30)	10.03 ± 1.34 (-0.94)	20.41 ± 1.41 (-1.81)	9.36 ± 1.15 (0.10)

Reconstruction Analysis

In Table 4, we evaluate the reconstruction quality for healthy brain regions. Overall, when comparing T1-weighted and T2-weighted images, T1-weighted images exhibit a slightly higher reconstruction error. Additionally, the external test sets (BraTS21, MSLUB, ATLAS, WMH) show increased reconstruction errors compared to the IXI test set. However, the overall performance trends are consistent with those observed in Table 1.

Table 4: Comparison of reconstruction quality of healthy structures across different models and datasets. The $l1$-error is computed using ground truth annotations of the healthy regions. For all metrics, the mean ± standard deviation across the different folds are reported. The arrow ↓ indicates that lower values are favorable. DDPM-based models are evaluated by ensembling different values for $t_{test} = [250, 500, 750]$

Model	***l*1-error** (e-3) ↓					
	IXI (T2)	BraTS21 (T2)	MSLUB (T2)	IXI (T1)	ATLAS (T1)	WMH (T1)
VAE	32.32 ± 0.64	31.52 ± 0.65	36.78 ± 0.79	39.47 ± 0.60	49.12 ± 0.84	51.22 ± 0.88
SVAE	29.08 ± 0.16	28.33 ± 0.10	33.21 ± 0.17	39.65 ± 0.50	51.51 ± 0.87	54.46 ± 1.14
AE	31.67 ± 0.41	30.95 ± 0.40	35.77 ± 0.34	39.04 ± 0.38	48.18 ± 0.50	50.08 ± 0.66
RA	34.36 ± 1.43	34.09 ± 1.71	37.89 ± 1.52	40.66 ± 2.46	48.25 ± 2.11	49.21 ± 1.88
PHANES	38.70 ± 1.74	35.19 ± 1.75	41.64 ± 2.05	44.03 ± 1.32	55.14 ± 1.23	58.31 ± 1.33
DAE	8.14 ± 0.17	11.52 ± 0.28	10.74 ± 0.14	9.84 ± 0.45	15.72 ± 0.52	14.92 ± 0.46
DDPM	14.29 ± 0.32	18.01 ± 1.10	16.54 ± 0.54	16.03 ± 0.17	23.4 ± 1.78	23.78 ± 1.89
pDDPM	9.70 ± 0.43	12.14 ± 0.32	11.52 ± 0.52	11.21 ± 0.28	19.03 ± 0.23	19.16 ± 0.36
cDDPM	9.68 ± 0.16	12.26 ± 0.17	11.54 ± 0.22	11.36 ± 0.44	16.83 ± 0.43	16.86 ± 0.47

MRI Scanner Details

Table 5: MRI scanner details and age statistics for various datasets used in this study. The table includes scanner model, field strength, echo time (TE), repetition time (TR), and flip angle values for different MRI sequences (T1 and T2). Some datasets do not have fully specified scanner parameters (missing information is denoted by -). * indicates that demographic information is only available for a subset of the data.

Dataset	**Age (mean/std)**	**Model**	**Field Strength [T]**	**Echo Time [ms]**	**Repetition Time [ms]**	**Flip Angle [°]**
IXI (T1/T2)	49.4/16.7	Philips Intera	1.5	9.6 / 5725.8	4.6 / 100.0	8.0 / 90.0
		Philips Gyroscan Intera	3.0	9.8 / 8178.3	4.6 / 100.0	8.0 / 90.0
		GE	1.5	-	-	-
BRATS (T2)	61.2/12.0*	-	-	-	-	-
MSLUB (T2)	39.3/10.1	Siemens Magnetom Trio MR	3.0	6000.0	120.0	120.0
ATLAS (T1)	-	-	1.5	-	-	-
		-	3.0	-	-	-
WMH (T1)	-	Philips Achieva	1.5	7.9	4.5	-
		Philips Ingenuity	3.0	9.9	4.6	-
		Siemens TrioTim	3.0	300	1.9	-
		GE Signa HDxtT	1.5	12.3	5.2	-
		GE Signa HDxt	3.0	7.8	3.0	-

Segmentation Maps

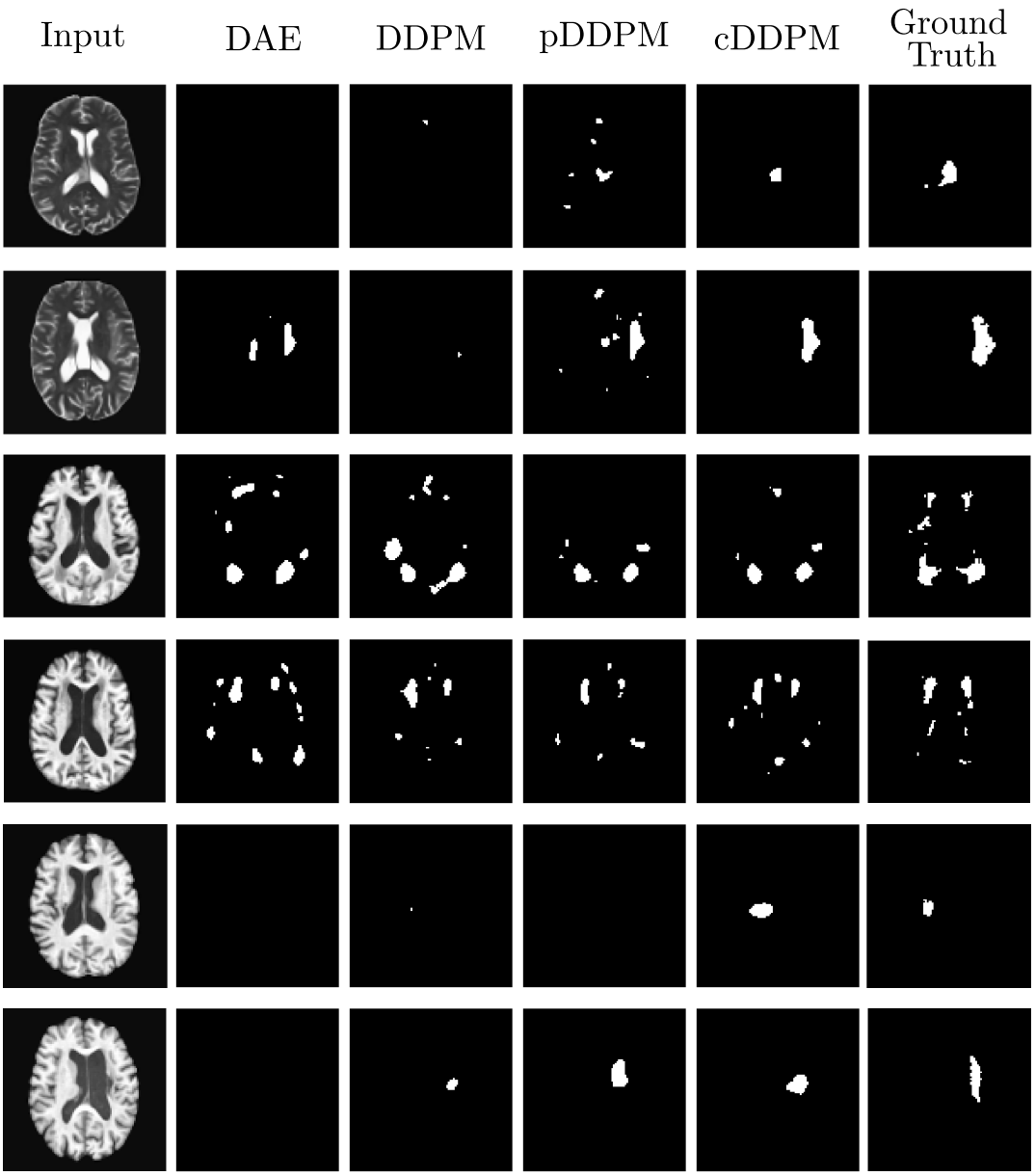


Figure 8: Segmentation maps corresponding to the residual maps in Figure 7, derived from the BraTS21 dataset (rows 1 and 2), the WMH dataset (rows 3 and 4), and the ATLAS dataset (rows 5 and 6). The binarization threshold for generating the segmentation maps was determined by optimizing for the highest possible Dice score.

8.4 Diffusion Models with ensembled Structure-based Anomaly Scoring for Unsupervised Anomaly Detection [23]

2024 IEEE International Symposium on Biomedical Imaging (ISBI) | 979-8-3503-1333-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISBI56570.2024.10635828

DIFFUSION MODELS WITH ENSEMBLED STRUCTURE-BASED ANOMALY SCORING FOR UNSUPERVISED ANOMALY DETECTION

Finn Behrendt[⋆] *Debayan Bhattacharya*[⋆] *Lennart Maack*[⋆] *Julia Krüger*[†]
Roland Opfer[†] *Robin Mieling*[⋆] *Alexander Schlaefer*[⋆]

[⋆] Institute of Medical Technology and Intelligent Systems,
Hamburg University of Technology, Hamburg, Germany
[†]Jung Diagnostics GmbH, Hamburg, Germany

ABSTRACT

Supervised deep learning techniques show promise in medical image analysis. However, they require comprehensive annotated data sets, which poses challenges, particularly for rare diseases. Consequently, unsupervised anomaly detection (UAD) emerges as a viable alternative for pathology segmentation, as only healthy data is required for training. However, recent UAD anomaly scoring functions often focus on intensity only and neglect structural differences, which impedes the segmentation performance. This work investigates the potential of Structural Similarity (SSIM) to bridge this gap. SSIM captures both intensity and structural disparities and can be advantageous over the classical $l1$ error. However, we show that there is more than one optimal kernel size for the SSIM calculation for different pathologies. Therefore, we investigate an adaptive ensembling strategy for various kernel sizes to offer a more pathology-agnostic scoring mechanism. We demonstrate that this ensembling strategy can enhance the performance of DMs and mitigate the sensitivity to different kernel sizes across varying pathologies, highlighting its promise for brain MRI anomaly detection.

Index Terms— Unsupervised Anomaly Detection, Diffusion Models, Brain MRI, SSIM

1. INTRODUCTION

While supervised deep learning techniques have demonstrated encouraging outcomes in detecting and segmenting anomalies in brain MRI scans [1], they depend on annotated and balanced data sets [2]. The need for such large data sets, especially with voxel-level annotations, is a challenge given the labor and resource-intensive requirements. Unsupervised anomaly detection (UAD) methods take a different approach and only require healthy or normal data for training. In reconstruction-based UAD, a generative model (GM) is trained to replicate MRI scans from a healthy training set. Following this, discrepancies between an MRI scan and the GM's replication can be used to identify abnormal patterns in unhealthy brain scans. Hence, unlike supervised models, UAD techniques are not dependent on pixel-level annotations of diseases and have the invaluable potential to identify any kind of irregularity that differs from a norm learned from the healthy training distribution.

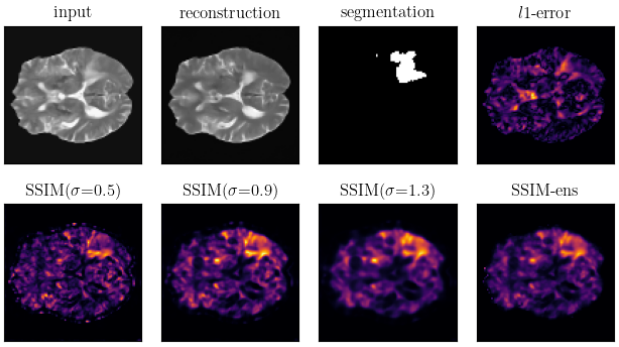


Fig. 1. Visualization of $l1$-based and SSIM-based anomaly scores for different values of σ and our ensemble solution SSIM-ens. Example is taken from the BraTS21 data set.

In the domain of UAD for brain MRI, various generative models (GMs) are used, with Autoencoders (AE) and their variational equivalents (VAE) being among the most common [3]. Recently, denoising diffusion probabilistic models (DDPM) have emerged as promising options, offering precise reconstructions, effective representation of the training set, and stable training characteristics [4, 5, 6]. While the type and architecture of the GM play a crucial role in UAD, another vital design parameter is the type of discrepancy measurement used to score anomalies. Predominantly, methods rely on the mean squared error ($l2$-error) or mean absolute error ($l1$-error). However, these metrics often miss structural differences and only focus on intensity-based discrepancies [7, 8]. Consequently, subtle anomalies of smaller intensity fluctuations might not be detected and reconstruction errors can be over-penalized. An alternative anomaly score that enables the detection of both intensity-based and structural discrepancies

is the Structural Similarity (SSIM) [9]. SSIM can provide a more balanced assessment, accounting for structural integrity, as demonstrated in Figure 1. Hence, existing literature suggests its application for UAD, either for industrial defect detection [10] or brain MRI pathology identification [8, 11]. Bergmann et al. [10] leverage a straightforward AE-based framework for detecting industrial defects, exchanging the $l2$-error with SSIM. Meissen et al. [8] derive the anomaly score by computing the SSIM on reconstructed AE features rather than in the pixel space.

In this work, we investigate possible synergies of SSIM with the strong reconstruction abilities of recent DDPMs. Our experimental results show that SSIM can improve the UAD performance when applied together with DDPMs. However, the kernel dimension σ introduces an additional hyperparameter that limits the generalization across different pathology types. Our findings underscore that this parameter plays a pivotal role in the detection of pathologies, with distinct pathologies of certain sizes showing preferences for specific kernel dimensions. To counteract this, we investigate utilizing an adaptive weighted average of multiple SSIM measurements over a spectrum of kernel sizes, mitigating the parameter sensitivity of SSIM. This method, which we call SSIM-ens, reduces the impact of individually chosen kernel sizes and leads to more robust performance in detecting a diverse range of pathologies, surpassing the results obtained with the traditional $l1$-error as anomaly score.

2. METHODS

We adopt the reconstruction-based UAD methodology, selecting DDPMs as our GM of choice. The UAD principle involves training the DDPM to reconstruct MRI scans without anomalies. During testing, discrepancies between the input MRI scan and its reconstruction are flagged. These disparities often hint at structures not encountered during the training phase, suggesting potential abnormalities.

2.1. Diffusion Models

DDPMs are generative models designed to capture the inherent data distribution of a given data set. In the training phase, DDPMs gradually transform an input image into Gaussian noise through a forward process, which is then reversed in the backward process to reconstruct the original image. This transformation is guided by a predefined schedule and involves adding controlled amounts of noise. The forward transformation is captured by: $\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0)$ where t determines the noise level that is sampled from $t \sim Uniform(1, ..., T)$ and $\boldsymbol{x}_0$ denotes a noise-free image. The backward process recovers the original image using: $\boldsymbol{x}_0^{rec} \sim p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$. Similar to [5], instead of predicting the added noise, we focus on directly estimating the reconstructed image, leading to the loss function: $\mathcal{L}_{rec} = |\boldsymbol{x}_0 - \boldsymbol{x}_0^{rec}|$. At test time, we directly estimate the reconstructed image based on the input, setting a fixed noise level t_{test}.

2.2. Anomaly Scoring with SSIM

SSIM is designed to measure the local similarities of two images where high values of SSIM reflect similar images and low values indicate differences across the two images. Instead of solely calculating intensity-based discrepancies, SSIM incorporates changes in structural information, contrast, and luminance. The SSIM equation for two images $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where μ_x and μ_y are the local means of x and y, respectively; σ_x^2 and σ_y^2 are the local variances; and σ_{xy} is the local covariance. The constants C_1 and C_2 serve to stabilize the division with a weak denominator. The local statistics required for these computations are derived using a Gaussian kernel, with the σ parameter defining the kernel's spread. The kernel dimension is derived by multiplying the spread with a constant factor $k_{dim} = int(3.5 * \sigma + 0.5) * 2 + 1$.

Adaptive Ensembling of SSIM (SSIM-ens)

Calculating SSIM is sensitive to the spread of the Gaussian Kernel where larger values for σ enlarge the considered neighborhood for a given pixel and vice versa. This can pose a challenge, as it affects the scale at which discrepancies are detected. To attenuate this dependency, we introduce SSIM-ens, an adaptive ensemble method that combines SSIM calculations over various σ values. This approach is intended to leverage the SSIM measurements at multiple scales, thereby providing an adaptive and more robust anomaly score. For a given input image $\boldsymbol{x}_0$ and its reconstruction $\boldsymbol{x}_0^{rec}$, the SSIM-ens anomaly score is calculated as a weighted average across different σ values $S = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$:

$$\text{SSIM-ens}(\boldsymbol{x}_0, \boldsymbol{x}_0^{rec}) = 1 - \sum_{i=1}^{n} \boldsymbol{w}_i \cdot \text{SSIM}_{\sigma_i}(\boldsymbol{x}_0, \boldsymbol{x}_0^{rec}),$$

with

$$\boldsymbol{w}_i = \frac{e^{-\text{SSIM}_{\sigma_i}(\boldsymbol{x}_0, \boldsymbol{x}_0^{rec})}}{\sum_{j=1}^{n} e^{-\text{SSIM}_{\sigma_j}(\boldsymbol{x}_0, \boldsymbol{x}_0^{rec})}}$$

where SSIM_{σ_i} is the individual SSIM calculated with the σ_i value, and $\boldsymbol{w}_i$ is a normalized exponential weighting factor inversely related to the corresponding SSIM_{σ_i} score, enhancing the focus on areas with higher discrepancies and potential anomalies. This results in a scoring mechanism that effectively captures variations in anomaly manifestation across different pathology types and sizes.

3. EXPERIMENTAL SETUP

3.1. Data Sets

We utilize the Information eXtraction from Images (IXI) data set that contains healthy brain MRI scans as the training set. The IXI data set includes 560 pairs of T1- and T2-weighted MRI scans. We sample a healthy test set (N_{test}=158) and partition the remaining samples into five healthy training sets (N_{train}=358) and 5 healthy validation sets (N_{val}=44) for five-fold cross-validation.

For evaluation, we rely on pathology data sets that provide pixel-wise annotations of pathology regions. Namely, we utilize the BraTS21, the ATLAS (v2), the WMH and MSLUB data sets, containing tumors, stroke, white matter hyperintensities and multiple sclerosis lesions, respectively. For the BraTS21 and MSLUB data sets, 1251 and 30 T2-weighted MRI scans and for the ATLAS and WMH data sets, 655 and 60 T1-weighted samples are available, respectively. For each evaluation data set, we split an unhealthy validation set of 100, 10, 175 and 15 samples for BraTS21, MSLUB, ATLAS and WMH respectively and use the remaining data as test set. Pre-processing includes resampling to an isotropic resolution of $1\,\mathrm{mm} \times 1\,\mathrm{mm} \times 1\,\mathrm{mm}$ and affine registration to the SRI24 Atlas. Subsequently, we skull-strip the brain scans and perform N4 bias field correction. Lastly, the volume resolution is reduced by a factor of two and the 15 top and bottom slices parallel to the transverse plane are removed, leading to a final resolution of $96 \times 96 \times 50$ voxels. For post-processing of the anomaly maps, median filtering with a kernel size of 5x5x5, brain mask eroding and connected components analysis is applied, similar to [3, 5]. For binarization, we calculate a threshold that shows the highest segmentation performance based on the unhealthy validation sets.

3.2. Implementation Details

We utilize a Unet, with channel dimensions of [128, 256, 256] as a denoising network[1]. We utilize structured simplex noise, which enhances the UAD efficacy of DDPMs in MRI images [4]. For the sampling process, we sample t uniformly from the interval $[0, 999]$ during training. At test time, we employ a consistent value for t by setting $t_{test} = 500$. For the SSIM-ens, we utilize a range of σ values $S = \{0.3, 0.5, ..., 1.5, 1.7\}$. We utilize the Adam optimization algorithm, with a learning rate of 1×10^{-5} and a batch size of 32. After training for 1600 epochs, model selection is based on the lowest reconstruction error achieved on the healthy validation set. Data processing is performed slice-by-slice; during training, slices are randomly selected with replacement, while at test time, we iterate through every slice to reconstruct the entire volume. In addition to the DDPMs, we implement AEs [3] and denoising AEs (DAE) [12] as baselines.

[1]Code available at `https://github.com/FinnBehrendt/Ensembled-SSIM-for-Unsupervised-Anomaly-Detection`

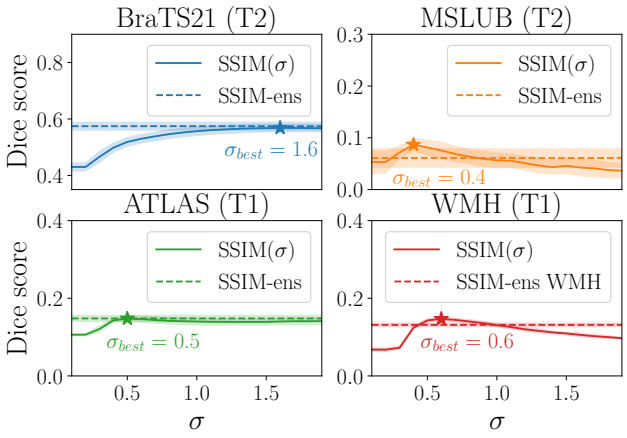


Fig. 2. Assessment of segmentation performance of DDPMs with SSIM using the Dice coefficient across varying SSIM parameter σ values. Optimal performance instances are denoted by stars (σ_{best}). The solid lines delineate the mean performance, while the surrounding shaded regions depict the standard deviation. The performance achieved by our SSIM-ens method is illustrated with dashed lines.

4. RESULTS

Table 1. Comparison of the evaluated models with color-coded changes within each model group. For all metrics, the mean ± standard deviation across the different folds are reported.

Model	BraTS21 (T2) **DICE** [%]	MSLUB (T2) **DICE** [%]	ATLAS (T1) **DICE** [%]	WMH (T1) **DICE** [%]
AE [3] (*l*1)	31.51±1.94	7.23±0.90	14.91±0.33	4.53±0.36
AE [3] (SSIM-ens)	↓25.88±0.43	↓3.57±0.70	↓8.23±0.13	↑ 6.53±0.20
DAE [12] (*l*1)	45.37±4.40	3.88±1.35	8.53±0.28	7.31±1.02
DAE [12] (SSIM-ens)	↑ 59.24±0.63	↓1.83±0.16	↑ 15.03±0.52	↑ 8.76±0.38
DDPM [4] (*l*1)	44.25±1.49	4.80±1.98	12.90±0.89	10.03±1.06
DDPM [4] (SSIM-ens)	↑ 57.44±1.40	↑ 6.10±1.78	↑ 14.81±0.79	↑ 13.16±0.44

Initially, we explore the impact of the σ parameter of SSIM on the segmentation performance of DDPMs for various pathologies. Following this, we evaluate and compare baseline models, both with and without the integration of our SSIM-ens strategy. To evaluate the segmentation performance, we report the pixel-wise Dice coefficient (DICE).

In Figure 2 we observe that the σ parameter affects the segmentation performance across all data sets. Selecting an optimal singular σ value is not straightforward due to the presence of multiple optimal points that vary between data sets. In contrast, by implementing an adaptive ensemble of diverse σ values within SSIM-ens, consistent performance is maintained across all data sets.

In Table 1, it is evident that applying the SSIM-based anomaly score leads to notable performance improvements for DDPMs across all data sets, reliably surpassing the $l1$-error.

5. DISCUSSION AND CONCLUSION

In this work, we provide an evaluation of the SSIM as an anomaly score for UAD in brain MRI. We demonstrate that the performance of SSIM is tied to the values of the kernel dimensions specified by σ, where the optimal values of σ vary for individual pathology types and sizes. Our study investigates an extension, SSIM-ens, which mitigates the parameter-dependence problem and offers a robust solution when integrated into DDPMs.

We show that our ensemble strategy consistently outperforms traditional $l1$ anomaly scores across varying pathologies if applied to DDPMs. By taking a weighted average of different SSIM anomaly scores across multiple parameter settings, SSIM-ens adds a degree of universality to the SSIM since the weighting is determined by the SSIM scores themselves, which means that the method adapts to the data it's evaluating. For discrepancies that are more prominent at certain scales, those discrepancies will be given more weight in the final score.

A notable observation is that SSIM does not consistently improve the performance of AEs and DAEs. For AEs the reason is seen in blurry reconstructions, failing to accurately mimic normal anatomy. On the other hand, DAEs, due to their inherent biases, often reconstruct pathologies different from tumor-like anomalies, as indicated in [5]. This impacts the effectiveness of SSIM-ens, which depends on precise representations of healthy anatomy in both the input and the reconstructed output.

Overall, our findings underscore the significance of the choice of anomaly score metric. It appears that the type of anomaly score utilized can have a substantial impact on UAD performance, potentially overshadowing the effects of changing the architecture of the generative model. We show that even though increasing performance across different data sets, SSIM alone adds a parameter dependence that prevents an optimal solution for differing pathologies. The proposed ensembling strategy can reduce the pathology-specific search for hyper-parameters and offers a more general solution for UAD in brain MRI.

Ethical approval: This research study was conducted using public data. Therefore, ethical approval was not required.

Funding: This work was partially funded by grant number KK5208101KS0 and by the Free and Hanseatic City of Hamburg (Interdisciplinary Graduate School).

6. REFERENCES

[1] Alexander Selvikvåg Lundervold and Arvid Lundervold, "An overview of deep learning in medical imaging focusing on mri," *Zeitschrift fur medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[2] Justin M. Johnson and Taghi M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[3] Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni, "Autoencoders for unsupervised anomaly segmentation in brain mr images: A comparative study," *Medical Image Analysis*, vol. 69, pp. 101952, 2021.

[4] Julian Wyatt, Adam Leach, Sebastian M. Schmon, and Chris G. Willcocks, "Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise," in *CVPR, NTIRE Workshop*, pp. 650–656.

[5] Finn Behrendt, Debayan Bhattacharya, Julia Krüger, Roland Opfer, and Alexander Schlaefer, "Patched diffusion models for unsupervised anomaly detection in brain mri," in *MIDL*, 2023.

[6] Finn Behrendt, Debayan Bhattacharya, Robin Mieling, Lennart Maack, Julia Krüger, Roland Opfer, and Alexander Schlaefer, "Guided reconstruction with conditioned diffusion models for unsupervised anomaly detection in brain mris," *arXiv preprint arXiv:2312.04215*, 2023.

[7] Felix Meissen, Benedikt Wiestler, Georgios Kaissis, and Daniel Rueckert, "On the pitfalls of using the residual error as anomaly score," in *MIDL*, 2022.

[8] Felix Meissen, Johannes Paetzold, Georgios Kaissis, and Daniel Rueckert, "Unsupervised anomaly localization with structural feature-autoencoders," *arXiv preprint arXiv:2208.10992*, 2022.

[9] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[10] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger, "Improving unsupervised defect segmentation by applying structural similarity to autoencoders," *arXiv preprint arXiv:1807.02011*, 2018.

[11] Ioannis Lagogiannis, Felix Meissen, Georgios Kaissis, and Daniel Rueckert, "Unsupervised pathology detection: A deep dive into the state of the art," *IEEE transactions on medical imaging*, vol. PP, 2023.

[12] Antanas Kascenas, Nicolas Pugeault, and Alison Q. O'Neil, "Denoising autoencoders for unsupervised anomaly detection in brain mri," in *MIDL*, 2022, vol. 172, pp. 653–664.

8.5 Leveraging the Mahalanobis Distance to enhance Unsupervised Brain MRI Anomaly Detection [24]

Reproduced with permission from Springer Nature: [Finn Behrendt, et al., "Leveraging the Mahalanobis Distance to Enhance Unsupervised Brain MRI Anomaly Detection," *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, 394–404, 2024].

Leveraging the Mahalanobis Distance to enhance Unsupervised Brain MRI Anomaly Detection

Finn Behrendt[1][0000−0001−7191−6508], Debayan Bhattacharya[1], Robin Mieling[1][0000−0003−0262−2519], Lennart Maack[1], Julia Krüger[2], Roland Opfer[2][0000−0002−9911−5478], and Alexander Schlaefer[1][0000−0001−9201−8854]

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

Abstract. Unsupervised Anomaly Detection (UAD) methods rely on healthy data distributions to identify anomalies as outliers. In brain MRI, a common approach is reconstruction-based UAD, where generative models reconstruct healthy brain MRIs, and anomalies are detected as deviations between input and reconstruction. However, this method is sensitive to imperfect reconstructions, leading to false positives that impede the segmentation. To address this limitation, we construct multiple reconstructions with probabilistic diffusion models. We then analyze the resulting distribution of these reconstructions using the Mahalanobis distance to identify anomalies as outliers. By leveraging information about normal variations and covariance of individual pixels within this distribution, we effectively refine anomaly scoring, leading to improved segmentation. Our experimental results demonstrate substantial performance improvements across various data sets. Specifically, compared to relying solely on single reconstructions, our approach achieves relative improvements of 15.9%, 35.4%, 48.0%, and 4.7% in terms of AUPRC for the BRATS21, ATLAS, MSLUB and WMH data sets, respectively.

Keywords: Unsupervised Anomaly Detection · Diffusion Models · Mahalanobis Distance

1 Introduction

Deep learning (DL) methods show promise in tasks like the segmentation of brain pathologies in magnetic resonance imaging (MRI) scans [19]. However, supervised DL methods require pixel-level annotations for training. This requirement becomes a challenge, particularly for screening tasks, where any pathology has to be detected even if not represented in the training data. Unsupervised Anomaly Detection (UAD) offers an alternative approach by learning the distribution of healthy data and identifying anomalies as outliers. A prevalent strategy is using reconstruction-based techniques [2]. These methods train generative models (GM) on a data set composed solely of healthy brain MRI scans. The underlying assumption is that the GMs will fail to reconstruct anomalies or pathological

structures not present in the training data set. Therefore, anomaly maps for segmenting abnormal structures can be derived from the deviations between input and reconstruction. However, a critical challenge UAD methods face lies in their high sensitivity to errors stemming from imperfect reconstructions [23, 21, 10]. As a result, even healthy structures exhibit deviations in the anomaly map. Therefore, discriminating deviations caused by genuine pathologies from those arising due to imperfect reconstructions becomes challenging, leading to false positives in the final segmentation. While deviations from imperfect reconstructions are inevitable, analyzing multiple reconstructions of the same input can offer valuable insights into the normal variations within the distribution of pseudo-healthy reconstructions, potentially simplifying the discrimination. These multiple reconstructions can be sampled using probabilistic GMs. Previous approaches have primarily focused on comparing the average reconstruction with the corresponding input image [3, 2]. However, these approaches ignore the valuable information in the variance and covariance of pixels across different reconstructions. The inter-pixel covariance across reconstructions quantifies the relationship between pixel values at different locations. It can be utilized to achieve a more balanced decision when measuring the distance of individual input pixels to the pseudo-healthy distribution of healthy pixels. Therefore, we propose using the Mahalanobis distance (MHD) [20] to measure the divergence of pixels in the input image from the pseudo-healthy distribution of pixels across multiple reconstructions. We employ denoising diffusion probabilistic models (DDPM) [12] to generate a pseudo-healthy reference distribution of reconstructions based on an individual input image. We then calculate the MHD between the input and the pseudo-healthy distribution to refine anomaly scoring. By considering the MHD in the pixel space with a full covariance matrix, we account for inter-pixel covariance. This enables capturing spatial information of neighboring pixels and long-range dependencies across pixels, such as symmetries in the reconstructions. Our results indicate that refining anomaly scoring by the MHD can substantially enhance the segmentation performance of conditioned DDPMs (cDDPMs), particularly when considering the inter-pixel covariance of the generated pseudo-healthy distributions. Compared to cDDPMs relying on single reconstructions, our approach leads to relative improvements of 15.9%, 35.4%, 48.0%, and 4.7% considering the AUPRC for the BRATS21, ATLAS, MSLUB and WMH data sets, respectively.

1.1 Recent Work

For most reconstruction-based approaches, AEs and VAEs are employed as GMs. While these architectures are conceptually simple and show promise in capturing the underlying distribution of healthy training data, their reconstructions tend to be blurry [2], substantially mitigating the segmentation performance. Therefore, many approaches aim to improve the reconstruction quality by focusing on spatial context [34] or utilizing 3D information [6]. Also, vector quantized VAEs and soft intro VAEs are applied to UAD in brain MRI [25, 7, 8]. Recent studies have indicated the effectiveness of DDPMs for UAD in brain MRI [24, 32, 4, 5].

Overall, GMs applied to the UAD task have shown promising progress. However, a crucial requirement for reconstruction-based UAD methods is to reconstruct healthy anatomy while avoiding the trivial replication of the input image. This necessitates the regularization of GMs, such as through a bottleneck in the latent space or additional regularization tasks like dropout [3] or denoising [13]. Consequently, imperfect reconstructions become inevitable. However, probabilistic GMs offer the appealing property of sampling multiple reconstructions. The assessment of multiple reconstructions could add valuable information for discriminating anomalies from imperfect reconstructions in the anomaly map. However, only a few studies have explored using VAEs or Bayesian AEs with Monte Carlo dropout to sample multiple reconstructions [3, 2]. These studies primarily concentrate on the mean of the generated reconstructions, which has not been shown to improve performance. Other approaches utilize uncertainty estimation to normalize the anomaly map by the estimated variance of individual pixels [29, 21, 10]. While this approach can improve the segmentation performance, it does not explicitly consider covariance across pixels. However, inter-pixel dependencies could provide valuable insights for anomaly scoring. Therefore, in this work, we focus on the inter-pixel dependencies and variations across different pseudo-healthy reconstructions and employ the MHD to measure the deviation of input pixels from the distribution of pixels in healthy reconstructions. While the MHD is commonly used for outlier detection, its typical application is at the sample level within some aggregated feature space for sample-level anomaly detection [16, 31]. Furthermore, Saase et al. [28] apply the MHD in the pixel space using a healthy data set as a reference distribution, suggesting that simple statistical methods can compete with deep learning models. However, individual brains in the training data exhibit substantial differences. As a result, relying solely on these general population-based distributions could lead to a mismatch between individual test cases and the reference distribution, potentially impeding the segmentation.

2 Methods

We propose utilizing DDPMs to construct pseudo-healthy distributions specific to each individual case during evaluation. Subsequently, these case-specific distributions are employed as a reference to compute the Mahalanobis distance (MHD) in the pixel space to refine anomaly scoring.

2.1 Generating Pseudo-healthy Distributions with DDPMs

DDPMs are specialized in learning the distribution of training images $\boldsymbol{x} \in \mathbb{R}^{H,W,C}$, where H, W, and C represent the height, width, and the number of channels, respectively. The training involves two primary processes: a forward process and a backward process. In the forward process, an image $\boldsymbol{x}_0$ is incrementally transformed into Gaussian noise, represented as $\boldsymbol{x}_T = \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

This transformation is guided by a predetermined noise schedule $[\beta_1, ..., \beta_T]$. The intermediate image states $\boldsymbol{x}_t$ are generated by

$$\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \quad \bar{\alpha}_t = \prod_{s=0}^{t}(1-\beta_t).$$

The noise level at each time step $t \in [1, ..., T]$, influences $\boldsymbol{x}_t$, which can vary from being the original image (at $t = 0$) to complete noise (at $t = T$). In the backward process, the reconstruction of the original image $\boldsymbol{x}_0^{rec}$ from the noisy state $\boldsymbol{x}_T$ is given by

$$\boldsymbol{x}_0^{rec} \sim p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \quad p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)).$$

Following [12], $\boldsymbol{\mu}_\theta$ is estimated using a Unet [27], and $\boldsymbol{\Sigma}(t)$ is set to $\frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t\mathbf{I}$. The training process entails minimizing the variational lower bound, which is approximated by the straightforward objective of predicting the added noise $\boldsymbol{\epsilon}\theta(\boldsymbol{x}t, t)$, as demonstrated in [12]. This yields the simplified loss function

$$\mathcal{L}_{simple} = ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)||^2.$$

In the context of reconstruction-based UAD, our objective is not to create new images from pure noise but to reconstruct healthy brain anatomy given an input image. Therefore, during testing, $\boldsymbol{x}_0^{rec}$ is estimated from $\boldsymbol{x}_t$, determining the extent of noise in $\boldsymbol{x}_t$ by $t_{test} < T$. To generate a distribution of multiple reconstructions, we sample N versions of $\boldsymbol{x}_t$ by repeatedly resampling the additional noise and reconstructing each noised image by the denoising network. As we train the model on healthy data, this leads to a pseudo-healthy distribution consisting of N different reconstructions of the given input image.

2.2 Anomaly Scoring using Pseudo-Healthy Distributions and Mahalanobis Distance

Our goal is to leverage the informative variations within the pseudo-healthy distribution of reconstructions.

Averaged Reconstructions Initially, we calculate the mean reconstruction from multiple pseudo-healthy samples, represented as: $\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i^{rec}$ Here, $\boldsymbol{x}_i^{rec}$ denotes the, i-th pseudo-healthy reconstruction, and N represents the total number of reconstructions. The anomaly score is defined as the inverted pixel-wise Structural similarity index measure (SSIM) between the input image $\boldsymbol{x}$ and the mean reconstruction $\boldsymbol{\mu}$:

$$S_{mean}(\boldsymbol{x}, \boldsymbol{\mu}) = 1 - SSIM(\boldsymbol{x}, \boldsymbol{\mu}). \tag{1}$$

Note that we use the pixel-wise SSIM as it has been shown to improve the anomaly scoring compared to intensity-based metrics [22, 15]

Mahalanobis Distance The MHD is a statistical measure, quantifying the distance of a sample point from a multivariate reference distribution, considering its covariance. Employing the pixels of an input image as sample points that are compared to the pseudo-healthy distribution of reconstructed pixels, we can capture the degree of deviation of each pixel in the input image from what is 'expected' in the distribution of pseudo-healthy reconstructions. First, we start by calculating the MHD with a diagonal covariance matrix $\boldsymbol{\Sigma}_{diag} = diag(\boldsymbol{\sigma}^2) \in \mathbb{R}^{H \cdot W \times H \cdot W}$, where $\boldsymbol{\sigma}^2$ is the variance of each pixel across the N reconstructions: $\boldsymbol{\sigma^2} = \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{x}_i^{rec} - \boldsymbol{\mu})^2$. Note that $\boldsymbol{x}$ and $\boldsymbol{\mu}$ are flattened to a dimension of $\mathbb{R}^{H \cdot W}$. This yields

$$MHD_{diag}(\boldsymbol{x}) = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}_{diag}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}. \tag{2}$$

This approach represents a standardization and allows for scaling the distance between input pixels and the mean reconstruction by the variance of individual pixels across different reconstructions. However, the diagonal covariance matrix does not consider covariance across different pixels. To capture inter-pixel correlations, we extend our analysis to utilize a full covariance matrix, calculated as $\boldsymbol{\Sigma}_{full} = \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{x}_i^{rec} - \boldsymbol{\mu})(\boldsymbol{x}_i^{rec} - \boldsymbol{\mu})^\top$ with dimension $\mathbb{R}^{H \cdot W \times H \cdot W}$, leading to

$$MHD_{full}(\boldsymbol{x}) = \sqrt{(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}_{full}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}. \tag{3}$$

After reshaping the MHD map to the input image shape, the final anomaly map is obtained by a per-pixel multiplication of the MHD map with the initial anomaly map for $S_{MHD} = S_{mean} \cdot MHD_{diag}$, and $S_{sMHD} = S_{mean} \cdot MHD_{full}$, respectively.

2.3 Data

Following the principle of reconstruction-based UAD, we utilize data sets without pathologies for training while evaluating data sets that contain annotated pathologies.

For training, we utilize the IXI data set [9], consisting of MRI scans in both T1- and T2-weighting. We split the training set into a healthy test set (N=160) and partition the remaining samples into 5 training sets (N=358) and 5 validation sets (N=44) for cross-validation.

For evaluation, we utilize four different data sets, namely the BRATS21 [1] (N=1152), MSLUB [17] (N=30), ATLAS [18] (N=655) and WMH [14] (N=60) data sets that exhibit tumors, multiple sclerosis, Stroke and white-matter lesions as pathologies, respectively. Note that while we train on both weightings separately, we evaluate BRATS and MSLUB on T2-weightings and ATLAS and WMH on T1-weightings. Pre-processing of the data includes resampling to a voxel dimension of $1 \times 1 \times 1$ mm, skull-stripping, registration to the SRI ATLAS and N4 bias-correction. Furthermore, we crop 15 top and bottom slices and reduce the dimension by a factor of 2, leading to a resolution of $192 \times 192 \times 50$

voxels. During training, we process the volumes slice-wise, with slices sampled with replacement. During evaluation, we iteratively reconstruct all slices to obtain the full volume.

2.4 Implementation Details

In this work, we build upon cDDPMs proposed in [5] as a probabilistic GM. Compared to DDPMs, cDDPMs utilize an additional feature representation of the input image to guide the denoising process. We follow the architectural design of [5] with a 3-layer Unet with channel dimensions [128, 128, 256] as a denoising network. We calculate the SSIM anomaly score with a Gaussian kernel with a standard deviation of 1. When calculating the MHD, we add a small regularization term (1e-5) to the diagonal entries of Σ_{full} To ensure numerical stability during inversion. Additionally, we apply a Gaussian filter to the MHD map with a standard deviation of 1. We compare established state-of-the-art baselines for UAD in brain MRI, including AE [2], VAE [2], DAE [13], DDPM [32], pDDPM [4] and cDDPM [5] as reconstruction-based approaches. Moreover, we compare RD [11] and FAE [22] as feature-based methods and the self-supervised approaches PII [30] and DRAEM [33]. Finally, we evaluate the covariance model (CM) of [28], where the MHD is calculated with the healthy training set as a reference distribution. For AEs and VAEs, we set the latent dimension to 128. For VAEs, $\beta_{KLD} = 0.001$ is chosen. We train for 1600 epochs, using the ADAM optimizer, a learning rate of 1e-4 and a batch size of 32. For all DDPM-based models, we utilize simplex noise as introduced on [32]. We uniformly sample noise levels $t \in [1, T]$ with $T = 1000$ during training and set the noise to $t_{test} = \frac{T}{2} = 500$ during evaluation. All models are implemented in PyTorch v1.10 and trained on an NVIDIA A6000 graphics card[3]. For evaluation, we utilize the best possible Dice-Coefficient ($\lceil \text{Dice} \rceil$) and the Area under the precision-recall curve (AUPRC). Additionally, we employ the permutation test from the MLXtend library [26]. This test involves 10,000 rounds of permutations and a significance level set at $\alpha = 5\%$ to assess statistical differences.

3 Results

We compare the segmentation performance of different variants of our approach to established state-of-the-art baselines. We average the metrics across the five folds and report the mean ± standard deviation. We initially tested different values for the number of reconstructions N in the range $N = [5, 10, ..., 30]$ and observed a moderate improvement in performance up to $N = 10$, after which performance plateaued. Therefore, to balance performance and inference time, we selected $N = 10$ reconstructions for each input image.

[3] Code available at `github.com/FinnBehrendt/Mahalanobis-Unsupervised-Anomaly-Detection`

Table 1. Segmentation performance regarding ⌈Dice⌉ and AUPRC. The highest values are shown in **bold**, where underlines denote statistical significance ($p < 0.05$). S_{mean} denotes the averaging of multiple reconstructions to derive the anomaly map. S_{MHD} and S_{sMHD} denote the use of the MHD either with a diagonal covariance matrix or with a full covariance matrix, respectively.

Model	BRATS		ATLAS		MSLUB		WMH	
	⌈DICE⌉	AUPRC	⌈DICE⌉	AUPRC	⌈DICE⌉	AUPRC	⌈DICE⌉	AUPRC
CM [28]	20.47 ± 0.22	14.03 ± 0.29	12.52 ± 0.47	9.31 ± 0.69	5.24 ± 0.27	2.59 ± 0.17	5.59 ± 0.09	2.70 ± 0.08
AE [2]	36.69 ± 0.20	33.58 ± 0.29	14.03 ± 0.27	11.68 ± 0.36	6.22 ± 0.05	3.55 ± 0.05	9.44 ± 0.26	5.60 ± 0.21
VAE [2]	36.04 ± 0.91	32.84 ± 1.07	14.48 ± 0.38	12.09 ± 0.41	6.33 ± 0.14	3.67 ± 0.11	9.52 ± 0.23	5.71 ± 0.23
FAE [22]	44.60 ± 2.17	43.75 ± 0.46	17.76 ± 0.16	13.97 ± 0.10	6.85 ± 0.65	4.02 ± 0.10	8.81 ± 0.38	4.97 ± 0.22
RD [11]	32.57 ± 0.15	27.13 ± 0.16	19.69 ± 0.26	15.65 ± 0.20	6.48 ± 0.20	3.66 ± 0.18	7.48 ± 0.10	4.22 ± 0.09
DAE [13]	62.93 ± 0.55	64.76 ± 0.79	19.42 ± 0.87	17.73 ± 0.88	8.35 ± 0.45	5.64 ± 0.37	11.14 ± 0.47	7.92 ± 0.55
DRAEM [33]	32.75 ± 3.63	26.38 ± 4.43	12.80 ± 1.94	9.63 ± 1.77	5.78 ± 2.29	2.66 ± 1.14	6.25 ± 1.89	3.23 ± 1.11
PII [30]	40.83 ± 2.18	36.49 ± 2.63	9.73 ± 1.89	7.26 ± 1.59	9.46 ± 0.43	5.21 ± 0.33	6.59 ± 1.87	3.49 ± 1.02
DDPM [32]	49.46 ± 1.56	47.57 ± 1.89	15.09 ± 0.64	11.85 ± 0.47	9.97 ± 0.64	6.03 ± 0.37	13.91 ± 0.37	9.15 ± 0.44
pDDPM [4]	54.26 ± 0.54	53.39 ± 0.70	18.83 ± 0.38	15.92 ± 0.44	10.37 ± 0.67	6.40 ± 0.51	15.31 ± 0.29	10.70 ± 0.21
cDDPM [5]	54.39 ± 0.70	54.31 ± 0.83	19.85 ± 0.90	16.99 ± 0.74	11.58 ± 0.35	7.76 ± 0.30	16.03 ± 0.88	12.15 ± 0.91
cDDPM S_{mean}	58.53 ± 0.48	59.14 ± 0.57	21.06 ± 1.09	18.17 ± 0.93	11.75 ± 0.44	7.75 ± 0.49	**17.09 ± 1.24**	13.15 ± 1.25
cDDPM S_{MHD}	58.47 ± 0.59	61.28 ± 0.63	20.34 ± 1.26	17.51 ± 1.23	12.25 ± 0.62	7.99 ± 0.69	16.82 ± 1.68	13.34 ± 1.90
cDDPM S_{sMHD}	**64.72 ± 0.52**	**68.55 ± 0.63**	**26.67 ± 1.61**	**24.61 ± 1.57**	**15.44 ± 0.85**	**11.47 ± 0.79**	16.65 ± 1.45	**13.77 ± 1.57**

The results are shown in Table 1 and Fig. 1. Comparing the baseline models, DAEs exhibit strong segmentation performance for the BRATS data set but are surpassed by cDDPMs for other pathologies in terms of Dice scores. Similarly, feature-based approaches (FAE and RD) perform well on individual data sets but struggle with generalization across all pathologies. Self-supervised approaches (PII and DRAEM) demonstrate poor performance across most data sets. Additionally, the CM method is consistently outperformed across all data sets. Overall, cDDPMs perform robustly across the evaluated data sets while enabling probabilistic sampling of multiple reconstructions. Hence, we consider cDDPMs to generate the pseudo-healthy distributions required for the MHD calculation. Our preliminary experiments indicate that other DDPM variants, such as the baseline DDPMs and pDDPMs, can also be utilized.

We find that averaging multiple reconstructions in cDDPMs (S_{mean}) enhances segmentation performance across most data sets compared to using a single reconstruction. In contrast to leveraging the MHD with a diagonal covariance matrix (S_{MHD}), utilizing the MHD with a full covariance matrix (S_{sMHD}) consistently demonstrates improved or competitive performance across all data sets. Notably, compared to the baseline cDDPMs, sampling multiple reconstructions and calculating the sMHD increases the processing time from 0.4 s to 4.9 s per volume. As illustrated in Fig. 1 (a), refining the anomaly map of cDDPMs by the sMHD leads to focused anomaly maps. Considering Fig. 1 (b), we observe non-zero correlations across the entire brain. Specifically, there exists a symmetric pattern regarding the tumor region with negative correlations in the left hemisphere and positive correlations in the right hemisphere. Exemplary anomaly maps for different models are provided in the supplementary material.

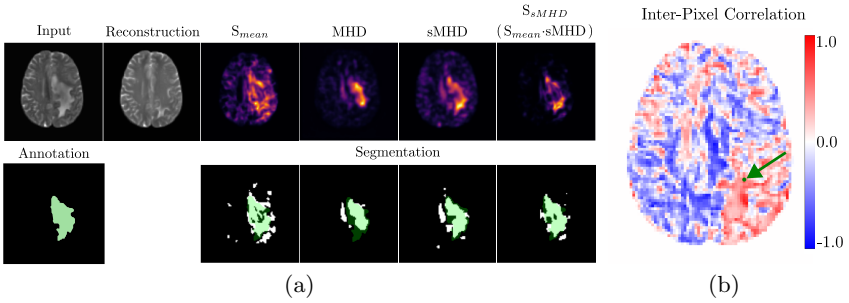


Fig. 1. (a): **Top row:** input, reconstruction, S_{mean} (SSIM), MHD, sMHD and the final anomaly map are shown for an exemplary image taken from the BRATS data set. **Bottom row:** the ground truth (green) and binarized segmentation maps (white) are shown. Note that the threshold for the segmentation maps is derived by optimizing the Dice score, based on the ground truth. (b): The correlation of one pixel (green arrow) with all other pixels, derived from Σ_{full} is visualized as a heatmap.

4 Discussion and Conclusion

A notable challenge of reconstruction-based UAD methods is their high sensitivity to imperfect reconstructions, often resulting in false positives that impede segmentation accuracy. To address this challenge, we propose to refine anomaly scoring by employing the MHD in the pixel space and identifying anomalies as outliers from pseudo-healthy distributions generated by cDDPMs.
Our results (as shown in Table 1) show that averaging multiple reconstructions from pseudo-healthy distributions (S_{mean}) can already improve segmentation performance. This improvement could be attributed to the variability of the noise structure added before reconstruction during the forward process of the cDDPM, resulting in regions with varying levels of complementary information available for denoising. Notably, applying the MHD with a diagonal covariance matrix (S_{MHD}) results in performance comparable to that of averaged reconstructions (S_{mean}). In contrast, using the spatial MHD (S_{sMHD}) substantially improves the segmentation performance. Fig. 1 (a) illustrates the differences between MHD and sMHD. It can be observed that the MHD is less sensitive to the edges of pathologies compared to sMHD. This indicates that the reconstructions exhibit higher variance in these regions, leading to a smaller weight in the anomaly map. In contrast, the anomaly map derived by sMHD shows improved pathology coverage. Fig. 1 (b) indicates the presence of inter-pixel correlations across the entire image, ranging from local neighborhoods to global dependencies exhibiting symmetry. However, the MHD with a diagonal covariance matrix does not capture these correlations. Consequently, the improved performance of sMHD compared to the MHD highlights the importance of considering these dependencies to identify abnormal pixels as outliers. Furthermore, our results indicate that considering the training data as a reference distribution for the MHD, as done in the case of CM [28], is ineffective for segmentation. This finding underscores the importance of constructing a pseudo-healthy reference distribution tailored to each individual test case, which is a key aspect of our approach.

In summary, leveraging the sMHD based on generated pseudo-healthy distributions for refining anomaly scoring can enhance the segmentation performance of DDPMs in the context of UAD in brain MRI. While we demonstrate this improved performance for cDDPMs, the baseline DDPMs and pDDPMs can also benefit from the sMHD, underscoring our method's versatility and potential impact in enhancing anomaly detection performance. A general limitation of UAD is its restriction to binary segmentation and the overall low performance for subtler anomalies, such as those found in WMH or MSLUB data sets. While our approach increases overall performance, it is important to acknowledge the increased computational demand due to the requirement of multiple reconstructions and matrix inversion. Future work could explore efficient approximations or decompositions to enhance the computational efficiency of MHD calculations.

Acknowledgments. This work was partially funded by grant number KK5208102HV3 and ZF4026303TS9 (Zentrales Innovationsprogramm Mittelstand) and by the Free and Hanseatic City of Hamburg (Interdisciplinary Graduate School).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
2. Baur, C., Stefan Denner, Benedikt Wiestler, Nassir Navab, Shadi Albarqouni: Autoencoders for unsupervised anomaly segmentation in brain mr images: A comparative study. Med. Image Anal. **69**, 101952 (2021). https://doi.org/10.1016/j.media.2020.101952
3. Baur, C., Wiestler, B., Albarqouni, S., Navab, N.: Bayesian skip-autoencoders for unsupervised hyperintense anomaly detection in high resolution brain mri. In: ISBI. pp. 1905–1909 (2020). https://doi.org/10.1109/ISBI45749.2020.9098686
4. Behrendt, F., Bhattacharya, D., Krüger, J., Opfer, R., Schlaefer, A.: Patched diffusion models for unsupervised anomaly detection in brain mri. In: MIDL (2023)
5. Behrendt, F., Bhattacharya, D., Mieling, R., Maack, L., Krüger, J., Opfer, R., Schlaefer, A.: Guided reconstruction with conditioned diffusion models for unsupervised anomaly detection in brain mris. arXiv preprint arXiv:2312.04215 (2023)
6. Bengs, M., Behrendt, F., Krüger, J., Opfer, R., Schlaefer, A.: Three-dimensional deep learning with spatial erasing for unsupervised anomaly segmentation in brain mri. IJCARS **16**(9), 1413–1423 (2021). https://doi.org/10.1007/s11548-021-02451-9
7. Bercea, C., Benedikt Wiestler, Daniel Rueckert, Julia A Schnabel: Generalizing unsupervised anomaly detection: Towards unbiased pathology screening. MIDL (2023)
8. Bercea, C.I., Wiestler, B., Rueckert, D., Schnabel, J.A.: Reversing the abnormal: Pseudo-healthy generative networks for anomaly detection. MICCAI **14224**, 293–303 (2023)

9. Biomedical Image Analysis Group: Ixi dataset – brain development, https://brain-development.org/ixi-dataset/
10. Cai, Y., Chen, H., Yang, X., Zhou, Y., Cheng, K.T.: Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images. Med. Image Anal. **86**, 102794 (2023). https://doi.org/10.1016/j.media.2023.102794
11. Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: CVPR. pp. 9737–9746 (June 2022)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
13. Kascenas, A., Pugeault, N., O'Neil, A.Q.: Denoising autoencoders for unsupervised anomaly detection in brain mri. In: MIDL (2022)
14. Kuijf, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., et al.: Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. IEEE TMI **38**(11), 2556–2568 (2019)
15. Lagogiannis, I., Meissen, F., Kaissis, G., Rueckert, D.: Unsupervised pathology detection: A deep dive into the state of the art. IEEE TMI **PP** (2023). https://doi.org/10.1109/TMI.2023.3298093
16. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. NeurIPS **31** (2018)
17. Lesjak, Ž., Galimzianova, A., Koren, A., Lukin, M., Pernuš, F., Likar, B., Špiclin, Ž.: A novel public mr image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. Neuroinformatics **16**(1), 51–63 (2018)
18. Liew, S.L., Lo, B.P., Donnelly, M.R., Zavaliangos-Petropulu, A., Jeong, J.N., Barisano, G., Hutton, A., Simon, J.P., Juliano, J.M., Suri, A., et al.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Scientific data **9**(1), 320 (2022)
19. Lundervold, A.S., Lundervold, A.: An overview of deep learning in medical imaging focusing on mri. Zeitschrift fur medizinische Physik **29**(2), 102–127 (2019). https://doi.org/10.1016/j.zemedi.2018.11.002
20. Mahalanobis, P.: On the generalised distance in statistics. In: Proceedings of the National Institute of Science of India. vol. 12, pp. 49–55 (1936)
21. Mao, Y., Xue, F.F., Wang, R., Zhang, J., Zheng, W.S., Liu, H.: Abnormality detection in chest x-ray images using uncertainty prediction autoencoders. In: MICCAI. pp. 529–538. Springer (2020)
22. Meissen, F., Paetzold, J., Kaissis, G., Rueckert, D.: Unsupervised anomaly localization with structural feature-autoencoders. arXiv preprint arXiv:2208.10992 (2022)
23. Meissen, F., Wiestler, B., Kaissis, G., Rueckert, D.: On the pitfalls of using the residual error as anomaly score. In: MIDL (2022)
24. Pinaya, W.H.L., Graham, M.S., Gray, R., Da Costa, P.F., Tudosiu, P.D., Wright, P., Mah, Y.H., MacKinnon, A.D., Teo, J.T., Jager, R., et al.: Fast unsupervised brain anomaly detection and segmentation with diffusion models. In: MICCAI (2022)
25. Pinaya, W.H.L., Tudosiu, P.D., Gray, R., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J.: Unsupervised brain imaging 3d anomaly detection and segmentation with transformers. Med. Image Anal. **79**, 102475 (2022). https://doi.org/10.1016/j.media.2022.102475
26. Raschka, S.: Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. The Journal of Open Source Software **3**(24) (2018)

27. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
28. Saase, V., Wenz, H., Ganslandt, T., Groden, C., Maros, M.E.: Simple statistical methods for unsupervised brain anomaly detection on mri are competitive to deep learning methods. arXiv preprint arXiv:2011.12735 (2020)
29. Sato, K., Hama, K., Matsubara, T., Uehara, K.: Predictable uncertainty-aware unsupervised deep anomaly segmentation. In: IJCNN. pp. 1–7. IEEE, Piscataway, NJ (2019). https://doi.org/10.1109/IJCNN.2019.8852144
30. Tan, J., Hou, B., Day, T., Simpson, J., Rueckert, D., Kainz, B.: Detecting outliers with poisson image interpolation. In: MICCAI (2021). https://doi.org/10.1007/978-3-030-87240-3–"textunderscore ″56
31. Vasiliuk, A., Frolova, D., Belyaev, M., Shirokikh, B.: Limitations of out-of-distribution detection in 3d medical image segmentation. JMI **9**(9) (2023). https://doi.org/10.3390/jimaging9090191
32. Wyatt, J., Leach, A., Schmon, S.M., Willcocks, C.G.: Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In: CVPR. pp. 650–656
33. Zavrtanik, V., Kristan, M., Skočaj, D.: Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8330–8339 (2021)
34. Zimmerer, D., Kohl, S., Petersen, J., Isensee, F., Maier-Hein, K.: Context-encoding variational autoencoder for unsupervised anomaly detection. In: MIDL (2019)

8.6 Combining Reconstruction-based Unsupervised Anomaly Detection with Supervised Segmentation for Brain MRIs [25]

Combining Reconstruction-based Unsupervised Anomaly Detection with Supervised Segmentation for Brain MRIs

Finn Behrendt[1] FINN.BEHRENDT@TUHH.DE
Debayan Bhattacharya[1] DEBAYAN.BHATTACHARYA@TUHH.DE
Lennart Maack [1] LENNART.MAACK@TUHH.DE
Julia Krüger[2] JULIA.KRUEGER@JUNG-DIAGNOSTICS.DE
Roland Opfer[2] ROLAND.OPFER@JUNG-DIAGNOSTICS.DE
Alexander Schlaefer[1] SCHLAEFER@TUHH.DE
[1] *Institute of Medical Technology and Intelligent Systems, Hamburg University of Technology, Hamburg, Germany*

[2] *Jung Diagnostics GmbH, Hamburg, Germany*

Abstract

In contrast to supervised deep learning approaches, unsupervised anomaly detection (UAD) methods can be trained with healthy data only and do not require pixel-level annotations, enabling the identification of unseen pathologies. While this is promising for clinical screening tasks, reconstruction-based UAD methods fall short in segmentation accuracy compared to supervised models. Therefore, self-supervised UAD approaches have been proposed to improve segmentation accuracy. Typically, synthetic anomalies are used to train a segmentation network in a supervised fashion. However, this approach does not effectively generalize to real pathologies. We propose a framework combining reconstruction-based and self-supervised UAD methods to improve both segmentation performance for known anomalies and generalization to unknown pathologies. The framework includes an unsupervised diffusion model trained on healthy data to produce pseudo-healthy reconstructions and a supervised Unet trained to delineate anomalies from deviations between input-reconstruction pairs. Besides the effective use of synthetic training data, this framework allows for weakly-supervised training with small annotated data sets, generalizing to unseen pathologies. Our results show that with our approach, utilizing annotated data sets during training can substantially improve the segmentation performance for in-domain data while maintaining the generalizability of reconstruction-based approaches to pathologies unseen during training.

Keywords: Unsupervised Anomaly Detection, Diffusion Models, Brain MRI, Self Supervision, Weak Supervision

1. Introduction

Deep learning (DL) methods have advanced in their ability to detect and segment brain pathologies in MRI images (Lundervold and Lundervold, 2019). However, acquiring annotated data for each pathology is a challenge, especially when considering screening tasks, where the objective is to detect any potential anomaly.
Unsupervised anomaly detection (UAD) provides a potential solution by modeling the distribution of healthy brain MRI scans to identify anomalies as outliers. A common technique in UAD is reconstruction-based anomaly detection, where generative models (GM) are trained to reconstruct healthy brain images. At test time, the GMs fail to replicate

pathologies, thereby revealing anomalies through discrepancies between input and reconstruction. This method only necessitates healthy data and enables the identification of pathologies not encountered during training, which poses a challenge for supervised models. However, the performance of reconstruction-based UAD methods is often surpassed by supervised models when sufficient task-specific data is available (Chen et al., 2020; Baur et al., 2021b). Unlike supervised methods, UAD methods that rely on reconstructions do not directly learn the relationship between abnormal patterns and their corresponding annotations. Instead, the segmentation map is a byproduct of measuring the discrepancy between input and reconstruction. This results in a noisy anomaly map with potential false positives caused by the GM's imperfect reconstructions. Consequently, distinguishing actual anomalies from normal reconstruction errors can be challenging. An alternative approach is self-supervised UAD, where synthetic anomalies are introduced to the healthy brain images to train a segmentation network in a supervised manner. Unlike reconstruction-based UAD, this strategy produces distinct anomaly maps with high specificity, simplifying the discrimination of abnormal structures similar to the synthesized anomalies. However, the segmentation performance depends on the nature of the generated anomalies and tends to have limited generalization to real pathologies (Lagogiannis et al., 2023; Cai et al., 2023).
In this study, we aim to combine the strong generalization capabilities and high sensitivity of reconstruction-based methods with the high specificity of self-supervised methods. We develop a framework that employs a denoising diffusion probabilistic model (DDPM; DM) to generate pseudo-healthy reconstructions of potentially abnormal input images (reconstruction branch). Furthermore, an Unet is trained to segment anomalies based on the residual of the input and the pseudo-healthy reconstruction (segmentation branch). We consider different settings to obtain the annotations for the supervised training of the Unet. First, in the self-supervised setting, we introduce synthetically generated anomalies to healthy brain MRIs. Second, in the semi-supervised setting, we utilize a small amount of annotated data containing real pathologies. At test time, the unsupervised anomaly maps from the reconstruction branch and the supervised predictions from the segmentation branch are fused to a final anomaly score.
The results demonstrate that in contrast to self-supervised methods, our approach allows to integrate supervision while maintaining the generalizability of the underlying reconstruction branch. Specifically, we can improve the Dice score of reconstruction-based UAD methods from 58.55 % to 69.68 % for tumors when using the same pathologies for training, while the Dice score for stroke lesions unseen during training increases from 24.74 % to 26.77 %.

2. Related Work

For reconstruction-based UAD, different architectures have been proposed as GM. While the majority focuses on Autoencoders (AE) (Baur et al., 2021a) or Variational autoencoders (VAE) (Zimmerer et al., 2019; Chen et al., 2020; Bercea et al., 2023a,c), also vector-quantized VAEs (Pinaya et al., 2022) and GANs (Nguyen et al., 2021) have been employed. Moreover, it has been shown that utilizing denoising tasks for regularization with Unet-like AEs can improve the UAD performance (Kascenas et al., 2022, 2023). Consequently, DDPMs have emerged as a GM for reconstruction-based UAD (Wyatt et al., 2022; Behrendt et al., 2023a,b; Bercea et al., 2023b). In self-supervised UAD, typically, synthetic anomalies

are incorporated into normal brain images. Subsequently, Unets are trained to segment these synthetic anomalies (Tan et al., 2021, 2022; Cho et al., 2022; Meissen et al., 2022a). We note that while AE-based reconstruction methods may also fall under the category of self-supervised techniques, within this work, the term "self-supervised" refers to the aforementioned approach of training segmentation models using synthetic anomalies. Expanding on this strategy, DRAEM (Zavrtanik et al., 2021) employs a dual-network architecture comprising a generator and a segmentation network. The generator is trained to eliminate synthetic anomalies, thereby providing a pseudo-healthy reconstruction. The segmentation network is then used to segment the generated anomalies, given the concatenation of abnormal input and pseudo-healthy reconstruction. Note that for the generator network in DRAEM, inpainting of synthetic anomalies is enforced by calculating the reconstruction loss between reconstruction and the anomaly-free input. In contrast, in our approach, the reconstruction model is trained on healthy data in an unsupervised fashion to remove any abnormal structure that is not part of the healthy training distribution. Hence, we expect this approach to generalize more readily to real pathologies. The authors (Liu et al., 2022) take a similar approach, aiming to improve supervised segmentation performance by augmenting a dual-branch Unet with pseudo-healthy reconstructions. These reconstructions are generated by a Soft-Intro VAE trained on healthy data. In contrast, our proposed framework does not solely depend on supervised predictions. Instead, these predictions are combined with the unsupervised anomaly scores derived from reconstructions of a DM. We hypothesize that this combination enables general anomaly detection, particularly for pathologies unseen during training.

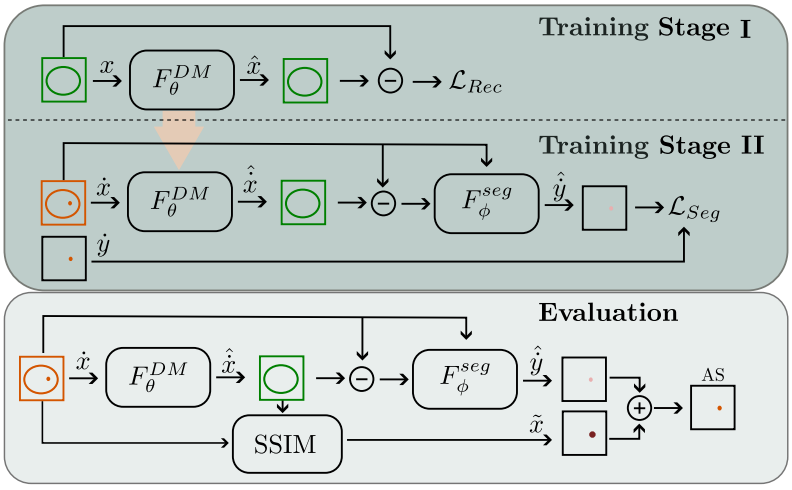


Figure 1: Schematic drawing of SADM. In Stage I, F_θ^{DM} is trained to reconstruct healthy brain images. In stage II, the parameters θ are fixed, and the segmentation network F_ϕ^{seg} is trained, either on synthetic anomalies (self-supervised) or real pathologies (semi-supervised). At test time, the supervised prediction $\hat{y}$ and the unsupervised anomaly map $\tilde{x}$ are combined to the final anomaly score (AS).

3. Method

In this section, we introduce our framework for supervised anomaly detection with DMs (SADM), detailed schematically in Figure 1.

3.1. Supervised Anomaly Detection with Diffusion Models (SADM)

SADM integrates two primary branches: a DM for generating pseudo-healthy reconstructions (reconstruction branch) and a supervised Unet for segmentation (segmentation branch). We train SADM in two sequential stages.

Stage I: Unsupervised Reconstruction

In the first stage, our objective is to train the DM to reconstruct healthy brain scans $\hat{\boldsymbol{x}} = F_\theta^{DM}(\boldsymbol{x})$ where $\boldsymbol{x} \in \mathbb{R}^{H\times W}$. The training of the DM focuses on optimizing parameters θ to minimize the $l1$-reconstruction loss:

$$\mathcal{L}_{Rec} = |\boldsymbol{x} - \hat{\boldsymbol{x}}|. \tag{1}$$

Stage II: Supervised Segmentation

In the second stage, the pseudo-healthy reconstruction generated by the DM trained in Stage I is utilized to support anomaly segmentation. Given an input scan with a real or synthetic anomaly $\dot{\boldsymbol{x}} \in \mathbb{R}^{H\times W}$ and its corresponding ground truth annotation $\dot{\boldsymbol{y}} \in \mathbb{R}^{H\times W}$, we use the DM, trained in Stage I, to generate the pseudo-healthy reconstruction $\hat{\dot{\boldsymbol{x}}} = F_\theta^{DM}(\dot{\boldsymbol{x}})$. Next, we feed both the residual $(\dot{\boldsymbol{x}} - \hat{\dot{\boldsymbol{x}}})$ and the original input $\dot{\boldsymbol{x}}$ into a Unet. After encoding both inputs, the resulting features are concatenated at each layer and fed to the Unet decoder to predict the segmentation map $\hat{\dot{\boldsymbol{y}}} = F_\phi^{seg}(\dot{\boldsymbol{x}} - \hat{\dot{\boldsymbol{x}}}, \dot{\boldsymbol{x}})$. Only the Unet parameters ϕ are optimized to minimize the cross-entropy (CE) segmentation loss during Stage II

$$\mathcal{L}_{Seg} = CE(\hat{\dot{\boldsymbol{y}}}, \dot{\boldsymbol{y}}). \tag{2}$$

Anomaly Detection

The anomaly detection process leverages both components of our framework for anomaly segmentation. Given a potentially abnormal input $\dot{\boldsymbol{x}}$, we generate a reconstruction $\hat{\dot{\boldsymbol{x}}} = F_\theta^{DM}(\dot{\boldsymbol{x}})$ by the DM. Next, we utilize F_ϕ^{seg} to derive the supervised anomaly prediction $\hat{\dot{\boldsymbol{y}}} = F_\phi^{seg}(\dot{\boldsymbol{x}} - \hat{\dot{\boldsymbol{x}}}, \dot{\boldsymbol{x}})$. In addition, we utilize the pixel-wise structural similarity (SSIM (Wang et al., 2004)) between input and reconstruction $\tilde{\boldsymbol{x}} = 1 - SSIM(\dot{\boldsymbol{x}} - \hat{\dot{\boldsymbol{x}}})$ for unsupervised anomaly scoring. The anomaly score (AS) is then derived as a combination of the unsupervised anomaly map and supervised anomaly prediction

$$\text{Anomaly Score (AS)} = \tilde{\boldsymbol{x}} + \hat{\dot{\boldsymbol{y}}}. \tag{3}$$

For pathologies similar to the anomalies seen during training, the supervised anomaly prediction will feature higher probabilities in abnormal regions, refining the unsupervised anomaly map. For unseen pathologies, the predicted probabilities are low such that $\tilde{\boldsymbol{x}}$ is unaltered. We hypothesize that this combination allows for comprehensive anomaly detection, leveraging the unsupervised anomaly map for general anomaly identification and the supervised prediction for precise segmentation of known abnormal patterns.

4. Experimental Setup

4.1. Data

We use T1-weighted MRIs from the IXI data set to train the DM in Stage I. We separate a healthy test set consisting of 160 samples. The remaining data is partitioned into five training sets (N=358) and validation sets (N=44) for cross-validation. In Stage II, we utilize the strategy applied in (Zavrtanik et al., 2021) to generate pairs of synthetic anomalies and ground truth annotation based on the IXI data set (DRAEM). Additional information about the generation process and exemplary anomalies are provided in Appendix C. Additionally, for the weakly supervised setting, we utilize small subsets containing approximately 10% of the BraTS21 (BRATS, N=1251) (Baid et al., 2021; Bakas et al., 2017; Menze et al., 2014), and ATLAS-v2.0 (ATLAS, N=655) (Liew et al., 2022) data sets. For evaluation, we utilize the remaining 1151 and 589 samples of the BRATS and ATLAS data sets, respectively. Furthermore, we utilize the augmented IXI test set (DRAEM) to assess the segmentation performance concerning synthetic anomalies.

Pre- and post-processing: We resample all T1 MRI scans to a resolution of $[1 \times 1 \times 1]$ mm and register them to the SRI24-Atlas (Rohlfing et al., 2010). Subsequently, we perform skull-stripping using HD-BET (Isensee et al., 2019) leading to volumes of size $[192 \times 192 \times 160]$ voxels. Finally, we apply bias-field corrections, reduce the resolution by a factor of two and crop 15 top and bottom slices in the transverse plane. For post-processing, we apply median filtering with a kernel size of 5 to the unsupervised anomaly maps.

4.2. Implementation Details

We utilize DMs as GM within our proposed framework to generate pseudo-healthy reconstructions[1]. Specifically, we use conditioned DDPMs (cDDPM) following the implementation of (Behrendt et al., 2023b). For the supervised segmentation of the residual image, we utilize a Unet (Ronneberger et al., 2015) like architecture, adapted from (Kascenas et al., 2022). The volumes are processed in a slice-wise fashion, sampling slices uniformly during training. At test time, we reconstruct the full volume by iterating over all slices. We compare our framework against different established baselines. We compare reconstruction-based AEs and VAEs (Baur et al., 2021a), FAEs (Meissen et al., 2022b), DDPMs (Wyatt et al., 2022), pDDPMs (Behrendt et al., 2023a) and cDDPMs (Behrendt et al., 2023b). Furthermore, we compare the feature-based reverse distillation method (RD) (Deng and Li, 2022), the self-supervised Poisson image interpolation (PII) (Tan et al., 2021) and DRAEM-Net (Zavrtanik et al., 2021) approaches. Note that for PII we perform the anomaly generation based on the IXI data set. For all reconstruction-based methods, we utilize SSIM for anomaly scoring with a Gaussian kernel with standard deviation of $\sigma_{ssim} = 1$, leading to a window size of $k_{ssim} = 9$. Implementation details of our proposed framework and compared baselines are provided in Appendix B.

1. Code available at https://github.com/FinnBehrendt/Supervised-Anomaly-Detection-with-Diffusion-Models

	Model	Training Data		Test Data					
				BRATS (real)		**ATLAS** (real)		**DRAEM** (synthetic)	
		$\mathcal{D}_{healthy}$	$\mathcal{D}_{unhealthy}$	[**DICE**]	**AUPRC**	[**DICE**]	**AUPRC**	[**DICE**]	**AUPRC**
I. Unsupervised	AE	IXI	None	39.16 ± 0.64	35.95 ± 0.70	14.14 ± 0.28	11.84 ± 0.37	9.91 ± 0.04	5.27 ± 0.04
	VAE	IXI	None	39.25 ± 0.50	36.07 ± 0.56	14.52 ± 0.37	12.18 ± 0.39	9.83 ± 0.14	5.28 ± 0.08
	DAE	IXI	None	55.93 ± 0.66	56.42 ± 0.84	19.95 ± 0.96	18.18 ± 0.98	12.50 ± 0.31	7.50 ± 0.22
	FAE	IXI	None	43.04 ± 0.49	42.04 ± 0.41	17.59 ± 0.15	13.91 ± 0.10	**19.60 ± 0.49**	**13.68 ± 0.25**
	RD	IXI	None	32.90 ± 0.65	28.31 ± 0.86	19.45 ± 0.25	15.51 ± 0.20	19.55 ± 0.60	13.17 ± 0.61
	DDPM	IXI	None	48.65 ± 0.90	46.93 ± 1.02	17.86 ± 0.87	14.70 ± 0.70	10.37 ± 0.23	6.04 ± 0.27
	pDDPM	IXI	None	55.93 ± 0.28	55.44 ± 0.36	21.79 ± 0.40	19.12 ± 0.43	14.59 ± 0.47	9.27 ± 0.31
	cDDPM	IXI	None	**58.55 ± 0.78**	**59.09 ± 0.91**	**24.74 ± 1.15**	**21.76 ± 0.98**	11.94 ± 0.52	7.31 ± 0.43
II. Self-Supervised	PII	None	PII	30.38 ± 2.46	24.66 ± 2.54	9.81 ± 1.93	7.31 ± 1.64	23.44 ± 1.61	15.09 ± 0.97
	DRAEM-Net	None	DRAEM	24.78 ± 4.21	18.49 ± 4.05	12.65 ± 1.90	9.51 ± 1.75	**79.77 ± 2.37**	**83.39 ± 2.34**
	Unet	None	DRAEM	40.75 ± 3.30	37.64 ± 3.92	16.91 ± 0.38	15.25 ± 0.26	76.03 ± 1.21	80.30 ± 1.32
	Unet_{res}	IXI	DRAEM	45.80 ± 3.22	44.05 ± 4.09	18.44 ± 0.47	16.81 ± 0.44	77.43 ± 1.16	81.93 ± 1.23
	SADM	IXI	DRAEM	50.81 ± 0.57	49.81 ± 0.81	23.82 ± 0.32	20.71 ± 0.35	73.77 ± 2.50	71.85 ± 3.02
	SADM_{res}	IXI	DRAEM	**60.53 ± 0.54**	**60.27 ± 1.02**	**27.78 ± 0.14**	**24.57 ± 0.13**	76.72 ± 1.30	75.45 ± 1.96
III. Weakly-Supervised	Unet	None	BRATS	64.81 ± 0.21	69.24 ± 0.33	11.82 ± 0.60	10.32 ± 0.61	**24.83 ± 1.10**	**20.96 ± 1.46**
	Unet_{res}	IXI	BRATS	67.01 ± 0.70	71.80 ± 0.87	17.33 ± 1.31	15.55 ± 1.50	19.93 ± 2.40	16.41 ± 2.64
	SADM	IXI	BRATS	69.01 ± 0.21	72.62 ± 0.46	25.25 ± 0.58	21.03 ± 0.50	14.93 ± 0.51	11.65 ± 0.66
	SADM_{res}	IXI	BRATS	**69.68 ± 0.48**	**73.34 ± 0.85**	**26.77 ± 0.65**	**23.22 ± 0.86**	17.11 ± 1.78	14.47 ± 1.91
	Unet	None	ATLAS	35.13 ± 2.97	32.87 ± 3.07	46.30 ± 0.72	46.37 ± 0.73	**29.11 ± 1.02**	**24.55 ± 1.91**
	Unet_{res}	IXI	ATLAS	36.82 ± 4.18	34.91 ± 4.92	47.36 ± 0.80	**47.61 ± 0.88**	22.07 ± 2.20	17.94 ± 2.39
	SADM	IXI	ATLAS	58.52 ± 0.60	57.17 ± 1.60	46.40 ± 0.17	44.71 ± 0.15	16.10 ± 1.10	12.81 ± 1.09
	SADM_{res}	IXI	ATLAS	**58.85 ± 0.44**	**57.68 ± 1.23**	**47.64 ± 1.40**	46.13 ± 1.36	17.77 ± 1.82	14.49 ± 1.73

Table 1: Segmentation performance regarding DICE and AUPRC. **Block I:** Unsupervised approaches, trained with healthy data. **Block II:**, Self-supervised approaches, trained with synthetic anomalies. **Block III:** Weakly-supervised approaches, trained with real pathologies. $\mathcal{D}_{healthy}$ and $\mathcal{D}_{unhealthy}$ represent the type of data used during training.

5. Experiments

For all our experiments, we evaluate the BRATS and ATLAS data sets containing real pathologies and the IXI data set augmented with synthetic anomalies (DRAEM). We report the mean ± standard deviation across the different folds for the best possible Dice Score ([DICE]) as well as the Area under Precision-Recall Curve (AUPRC) to assess the segmentation performance. We evaluate different variants of SADM. In SADM_{res}, the residual of input and reconstruction and the (abnormal) input are fed to the Unet, whereas in SADM, only the input is used. Furthermore, we consider Unet and Unet_{res}, where, in contrast to SADM only the prediction of the Unet is used, ignoring the anomaly map of the unsupervised reconstruction branch. In Appendix D, we provide an ablation study on the weighted combination of the segmentation and reconstruction branch.

5.1. Training with Synthetic Anomalies

We evaluate our approach in different settings. First, we assume the typical UAD case where only data with healthy labels is available. We use synthetic anomalies to obtain a supervised signal for the segmentation branch in SADM. We utilize the generation process proposed in DRAEM (Zavrtanik et al., 2021) to generate the anomalies. In this setting, we compare our framework to various UAD baselines. Results are reported in block I and block II of Table 1. Across the compared UAD baselines in block I, cDDPMs show the highest

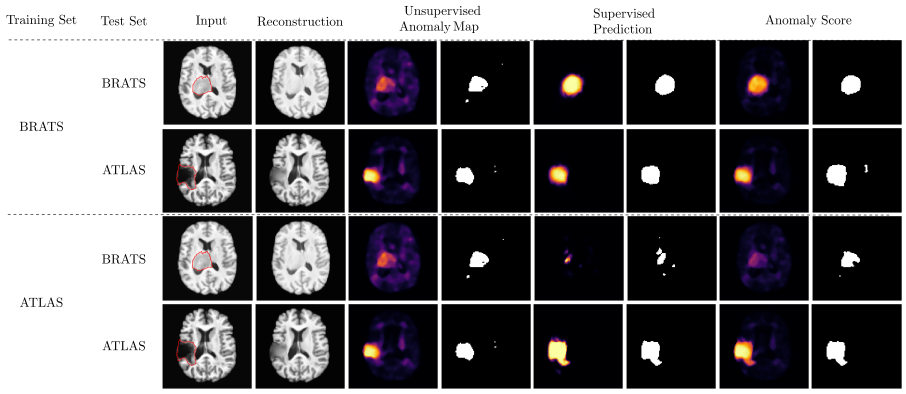


Figure 2: Examplary test cases for SADM$_{res}$, trained and evaluated in the weakly-supervised setting with the BRATS and ATLAS data sets, respectively. For visualization purposes, we provide exemplary binary segmentation maps for the unsupervised anomaly score, the supervised prediction and the final AS, respectively. We derive the binarization threshold by optimizing for the best possible dice score.

segmentation performance for real pathologies. Hence, we consider them as a reconstruction model for the SADM framework. For real pathologies, SADM$_{res}$ outperforms cDDPMs with performance improvements of 3.4 %, 12.3 % for the BRATS and ATLAS data sets, respectively. Considering the synthetic anomalies in the DRAEM data set, a substantially higher DICE of 76.72 % is reported for SADM$_{res}$ compared to the DICE of 11.94 % achieved by cDDPMs. Notably, while the DRAEM-Net shows relative performance improvements of 10.5 % over SADM$_{res}$ for synthetic anomalies, it fails to generalize to the real pathologies in the BRATS and ATLAS data sets. Even the Unet, trained with the same synthetic anomalies as in DRAEM-Net, outperforms DRAEM-Net considering real pathologies.
Comparing SADM and SADM$_{res}$, we observe that utilizing the residual of abnormal input and pseudo-healthy reconstruction in addition to the abnormal input substantially improves the segmentation performance across all data sets.

5.2. Training with Real Pathologies

In this section we investigate using our framework in a weakly-supervised setting. Instead of generating synthetic anomalies, we assume a small amount of annotated data is available and consider a subset of the BRATS and ATLAS data sets for training, respectively. We only train with one data set at a time to evaluate the generalization to unseen pathologies. The results for this weakly-supervised setting are reported in block III of Table 1. Using a small subset of annotated data substantially improves the segmentation of all models when evaluating the same (in-domain) data set. However, the segmentation performance of Unet and Unet$_{res}$ is poor for data sets containing pathologies unseen during training. In contrast, both SADM and SADM$_{res}$ enhance the segmentation performance on in-domain data while maintaining or even improving the performance of unsupervised cDDPMs for

unseen pathologies. A visualization of the anomaly maps coming from different branches of the SADM framework is provided in Figure 2.

6. Discussion and Conclusion

A significant challenge of supervised methods that UAD addresses is the need for annotated training data. This is especially crucial when considering screening tasks where the type and shape of potential lesions are unknown. Therefore, it is highly desirable to achieve generalization to different kinds of lesions while minimizing false positive predictions. In this work, we aim for a framework that benefits from the robust generalization of reconstruction-based UAD methods and the high discriminative power of supervised strategies.
Comparing the unsupervised and self-supervised approaches in Table 1, the additional shape information typically improves the segmentation performance with the magnitude of improvement dependent on the lesion type. However, considering purely self-supervised models, it is evident that supervised training based on synthetic data can result in overfitting. In contrast, our proposed framework, improves the segmentation performance for anomalies of known shape and appearance while maintaining or even improving the generalization of reconstruction-based UAD for pathologies unseen during training. This indicates that the framework effectively utilizes the complementary information of the reconstruction and segmentation branches, as highlighted in Figure 2. On the one hand, the supervised segmentation branch enhances the specificity for pathologies similar to the anomalies seen during training. On the other hand, the reconstruction branch maintains the high sensitivity of reconstruction-based UAD for any abnormal pattern unseen during the training of the DM. Furthermore, feeding the residual of input and reconstruction to the Unet in addition to the abnormal input can enhance the segmentation performance, particularly in the self-supervised setting. This indicates that the additional information in the residual may contribute to learning the deviation from a normal representation, potentially reducing the risk of overfitting to specific anomaly shapes. While the DRAEM-Net shares some similarities with our approach, there are significant differences. First, DRAEM-Net uses a generator network trained to remove synthesized anomalies. In contrast, our reconstruction branch employs a DM trained to reconstruct healthy data without explicitly enforcing the removal of specific anomalies. Second, instead of solely relying on the segmentation branch, we combine the supervised prediction with the unsupervised anomaly map derived from the reconstruction branch. As demonstrated in our experiments, these adaptations lead to improved segmentation performance and generalization, enabling the effective use of SADM in a weakly-supervised setting. Therefore, our framework adds a significant feature to UAD approaches, especially considering that some annotated data is typically available.
In summary, our approach shows encouraging results, paving the way for a practical solution for UAD in brain MRI. Limitations are seen in the potential reconstruction of unhealthy structures by the reconstruction branch and in the investigated synthetic anomalies intended initially for industrial defect detection. Despite the demonstrated improvement in performance, we anticipate further enhancements when integrating more realistic synthetic anomalies. Additionally, we intend to include data sets featuring subtler anomalies or different imaging modalities to broaden the evaluation of our approach.

Acknowledgments

This work was partially funded by grant number KK5208102HV3 and ZF4026303TS9 (Zentrales Innovationsprogramm Mittelstand) and by the Free and Hanseatic City of Hamburg (Interdisciplinary Graduate School).

References

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

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(1):1–13, 2017.

Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain mr images: A comparative study. *Medical Image Analysis*, 69:101952, 2021a. ISSN 1361-8415. doi: 10.1016/j.media.2020.101952. URL https://www.sciencedirect.com/science/article/pii/S1361841520303169.

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, 2021b. doi: 10.1148/ryai.2021190169.

Finn Behrendt, Debayan Bhattacharya, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Patched diffusion models for unsupervised anomaly detection in brain mri. In *Medical Imaging with Deep Learning*, 2023a. URL https://openreview.net/forum?id=0-uZr5S1tJE.

Finn Behrendt, Debayan Bhattacharya, Robin Mieling, Lennart Maack, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Guided reconstruction with conditioned diffusion models for unsupervised anomaly detection in brain mris. *arXiv preprint arXiv:2312.04215*, 2023b.

Cosmin Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A Schnabel. Generalizing unsupervised anomaly detection: Towards unbiased pathology screening. *Medical Imaging with Deep Learning*, 2023a. ISSN 2640-3498. URL https://openreview.net/forum?id=8ojx-Ld3yjR.

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

Cosmin I. Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A. Schnabel. Reversing the abnormal: Pseudo-healthy generative networks for anomaly detection. volume 14224, pages 293–303. 2023c. doi: 10.1007/978-3-031-43904-9_29.

Yu Cai, Hao Chen, Xin Yang, Yu Zhou, and Kwang-Ting Cheng. Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images. *Medical Image Analysis*, 86:102794, 2023. ISSN 1361-8415. doi: 10.1016/j.media.2023.102794.

Xiaoran Chen, Suhang You, Kerem Can Tezcan, and Ender Konukoglu. Unsupervised lesion detection via image restoration with a normative prior. *Medical Image Analysis*, 64:101713, 2020. ISSN 1361-8415. doi: 10.1016/j.media.2020.101713.

Jihoon Cho, Inha Kang, and Jinah Park. Self-supervised 3d out-of-distribution detection via pseudoanomaly generation. pages 95–103. Springer, Cham, 2022. doi: 10.1007/978-3-030-97281-3{\textunderscore}15. URL https://link.springer.com/chapter/10.1007/978-3-030-97281-3_15.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9737–9746, June 2022.

Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, Martin Bendszus, Klaus H. Maier-Hein, and Philipp Kickingereder. Automated brain extraction of multisequence mri using artificial neural networks. *Human brain mapping*, 40(17): 4952–4964, 2019. ISSN 1097-0193. doi: 10.1002/hbm.24750.

Antanas Kascenas, Nicolas Pugeault, and Alison Q. O'Neil. Denoising autoencoders for unsupervised anomaly detection in brain mri. In Ender Konukoglu, Bjoern Menze, Archana Venkataraman, Christian Baumgartner, Qi Dou, and Shadi Albarqouni, editors, *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, volume 172 of *Proceedings of Machine Learning Research*, pages 653–664. PMLR, 2022. URL https://proceedings.mlr.press/v172/kascenas22a.html.

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. doi: 10.1016/j.media.2023.102963.

Ioannis Lagogiannis, Felix Meissen, Georgios Kaissis, and Daniel Rueckert. Unsupervised pathology detection: A deep dive into the state of the art. *IEEE transactions on medical imaging*, PP, 2023. ISSN 0278-0062. doi: 10.1109/TMI.2023.3298093.

Sook-Lei Liew, Bethany P. Lo, Miranda R. Donnelly, Artemis Zavaliangos-Petropulu, Jessica N. Jeong, Giuseppe Barisano, Alexandre Hutton, Julia P. Simon, Julia M. Juliano, Anisha Suri, et al. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific data*, 9(1):320, 2022.

Huabing Liu, Dong Nie, Dinggang Shen, Jinda Wang, and Zhenyu Tang. Multimodal brain tumor segmentation using contrastive learning based feature comparison with monomodal normal brain images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 118–127. Springer, 2022.

Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift fur medizinische Physik*, 29(2):102–127, 2019. doi: 10.1016/j.zemedi.2018.11.002.

Felix Meissen, Georgios Kaissis, and Daniel Rueckert. Autoseg - steering the inductive biases for automatic pathology segmentation. pages 127–135. Springer, Cham, 2022a. doi: 10.1007/978-3-030-97281-3{\textunderscore}19. URL https://link.springer.com/chapter/10.1007/978-3-030-97281-3_19.

Felix Meissen, Johannes Paetzold, Georgios Kaissis, and Daniel Rueckert. Unsupervised anomaly localization with structural feature-autoencoders. *arXiv preprint arXiv:2208.10992*, 2022b.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. ISSN 0278-0062.

Bao Nguyen, Adam Feldman, Sarath Bethapudi, Andrew Jennings, and Chris G. Willcocks. Unsupervised region-based anomaly detection in brain mri with adversarial image inpainting. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1127–1131, 2021. doi: 10.1109/ISBI48211.2021.9434115.

Ken Perlin. An image synthesizer. *ACM Siggraph Computer Graphics*, 19(3):287–296, 1985.

Walter H. L. Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Unsupervised brain imaging 3d anomaly detection and segmentation with transformers. *Medical Image Analysis*, 79:102475, 2022. ISSN 1361-8415. doi: 10.1016/j.media.2022.102475.

Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. Torchio: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, 208:106236, 2021. ISSN 0169-2607.

Torsten Rohlfing, Natalie M. Zahr, Edith V. Sullivan, and Adolf Pfefferbaum. The sri24 multichannel atlas of normal adult human brain structure. *Human brain mapping*, 31(5): 798–819, 2010. ISSN 1097-0193. doi: 10.1002/hbm.20906.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention*, pages 234–241, 2015.

Jeremy Tan, Benjamin Hou, Thomas Day, John Simpson, Daniel Rueckert, and Bernhard Kainz. Detecting outliers with poisson image interpolation. pages 581–591. Springer, Cham, 2021. doi: 10.1007/978-3-030-87240-3{\textunderscore}56. URL https://link.springer.com/chapter/10.1007/978-3-030-87240-3_56.

Jeremy Tan, Benjamin Hou, James Batten, Huaqi Qiu, and Bernhard Kainz. Detecting outliers with foreign patch interpolation. *Machine Learning for Biomedical Imaging*, 1 (April 2022 issue):1–27, 2022. ISSN 2766-905X. URL https://www.melba-journal.org/papers/2022:013.html.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

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

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021.

David Zimmerer, Fabian Isensee, Jens Petersen, Simon Kohl, and Klaus Maier-Hein. Unsupervised anomaly localization using variational auto-encoders. In Shen, Dinggang and Liu, Tianming and Peters, Terry M. and Staib, Lawrence H. and Essert, Caroline and Zhou, Sean and Yap, Pew-Thian and Khan, Ali, editor, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 289–297, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32251-9.

Appendix A. Qualitative Comparison

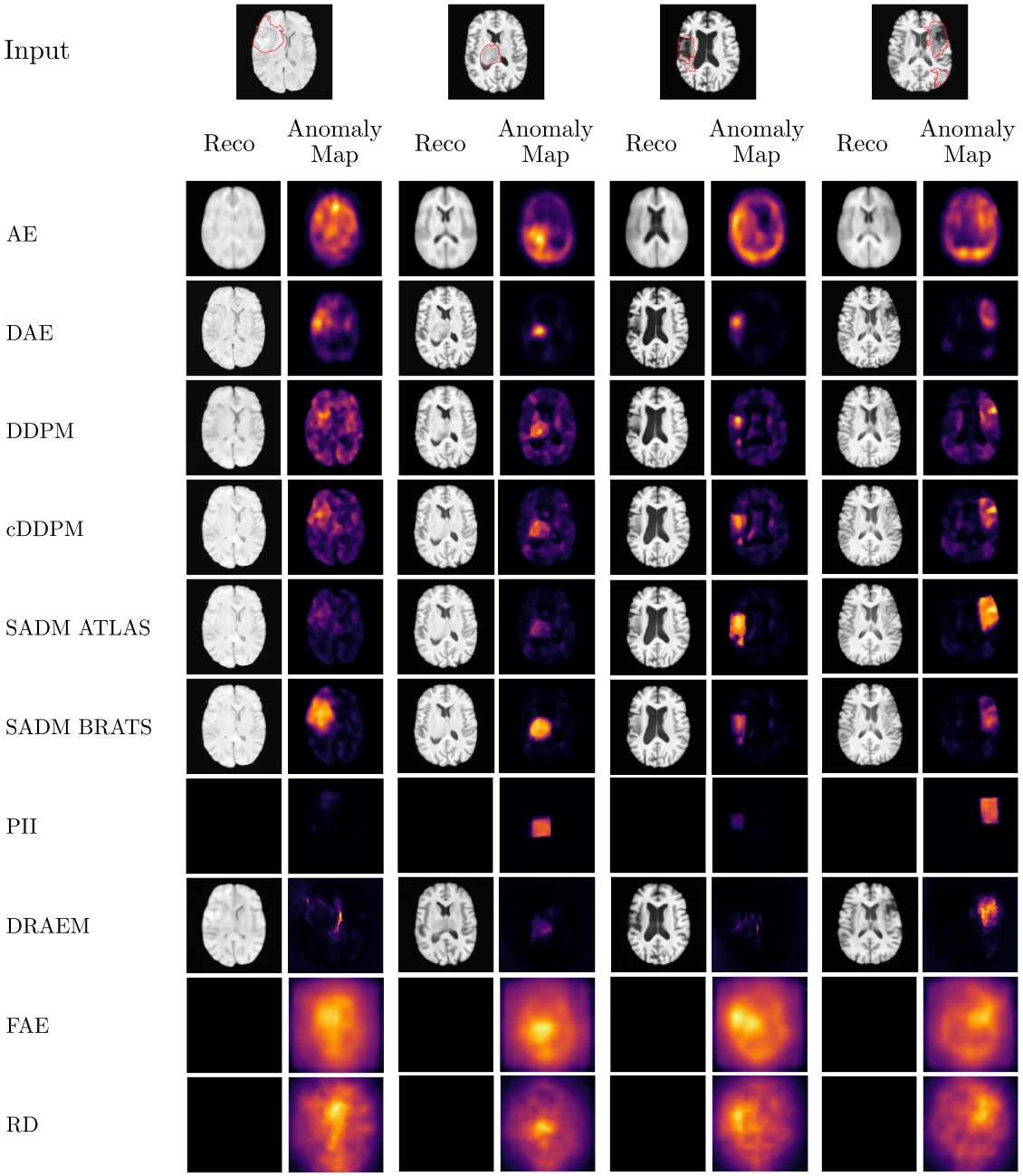


Figure 3: Comparison of baseline models for pathologies from the BRATS (left two columns) and ATLAS (right two columns) data sets.

Appendix B. Implementation Details

All models are implemented in Pytorch (v0.10). For data handling and augmentation, torchio (Pérez-García et al., 2021) is utilized. We choose the best-performing model checkpoint, measured by the validation set performance. We utilize Adam as an optimizer with a batch size of 32. For data augmentation, we utilize random -blur, -bias, -gamma and -ghosting. All Baselines are implemented following the official GitHub repositories. We train our models on NVIDIA RTX 3090 and V100 GPUs.

B.1. SADM

Our SADM framework consists of a reconstruction branch and a segmentation branch. In the reconstruction branch, we utilize cDDPMs (Behrendt et al., 2023b) as a generative model. We follow the official implementation[2] and utilize a 3-layer Unet with channel dimensions [128, 128, 256] as a denoising network with a pre-trained resnet50 encoder for conditioning. During training, we uniformly sample noise levels $t \in [0, T]$. At test time, we derive the final reconstruction as an average from reconstructions of different noise levels $t_{test} \in [250, 500, 750]$. For the segmentation branch, we adapt the Unet architecture as employed by (Kascenas et al., 2022). Our base Unet architecture consists of three layers with channel dimensions of 64, 128, and 256, respectively, incorporating group normalization and SiLU activation functions. For SADM$_{res}$, we utilize the same encoder to separately encode the residual of the input and reconstruction, as well as the input itself. The resulting feature maps are then concatenated along the channel dimension at each layer and passed to the decoder, effectively doubling the channel dimensions. A sigmoid layer is added after the final convolution to produce the segmentation output. In stage I and II, we train for 1600 and 600 epochs, with learning rates of 1e-4 and 5e-5, respectively.

B.2. Baselines

We implement various baseline methods based on the official code with individual adaptations of hyper-parameters that have been shown to improve training stability or performance regarding the validation data. Unless stated otherwise, all models are trained for 1600 epochs, choosing the best checkpoint based on the validation set performance, using Adam as an optimizer. For AEs and VAEs, we use a latent dimension of 128 and set the learning rate to 1e-4. For VAEs, we set $\beta_{KLD} = 0.001$. For RD and DRAEM, we set the learning rate to 1e-4. The DDPM, pDDPM and cDDPM baselines are trained with simplex noise as proposed in (Wyatt et al., 2022) and a learning rate of 1e-5, respectively. Note that for all DDPM-based baselines, we utilize the averaged reconstruction from three different noise levels $t_{test} \in [250, 500, 750]$.

Appendix C. Synthetic Anomalies

We generate the synthetic anomalies by following the procedure of (Zavrtanik et al., 2021). First, a noise image is generated using Perlin noise (Perlin, 1985), capturing a wide variety of shapes. Subsequently, the noise image is binarized by a uniformly sampled threshold,

2. https://github.com/FinnBehrendt/Conditioned-Diffusion-Models-UAD

resulting in an anomaly map M_a, that is used as ground truth annotation. The binary map is further processed by three random augmentation functions, sampled from the set of {posterize, sharpness, solarize, equalize, brightness change, color change, auto-contrast}, leading to I_{aug}. Finally, I_{aug} is masked by M_a and blended with the original image I, leading to $I_{syn} = (1 - M_a) \odot I + (1 - \gamma)(M_a \odot I) + \gamma(M_a \odot I_{aug})$. The operator $\odot$ denotes element-wise multiplication and γ denotes the opacity parameter that is uniformly sampled from $\gamma \in [0.2, 1.0]$. Figure 4 showcases exemplary synthetic images with the corresponding annotation mask.

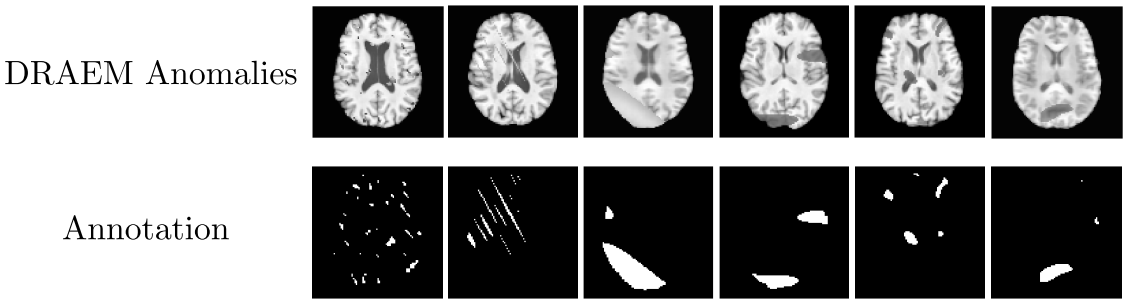


Figure 4: Examplary Synthetic Anomalies generated by the DRAEM procedure. Top: Images from the IXI data set, augmented with synthetic anomalies. Bottom: Annotation corresponding to the introduced anomalies.

Appendix D. Analysis of the Anomaly Score Weighting

In this section, we analyze the different weightings of the anomaly scores from the supervised and reconstruction branches. We derive the AS by weighing the individual scores as follows

$$\text{Anomaly Score (AS)} = \beta \cdot x_{tilde} + (1 - \beta) \cdot \dot{\hat{y}}. \tag{4}$$

We vary the weighting parameter β from zero to one. $\beta = 0$ corresponds to solely relying on the supervised branch (Unet_{res}). $\beta = 1$ corresponds to solely using the reconstruction branch (cDDPM).

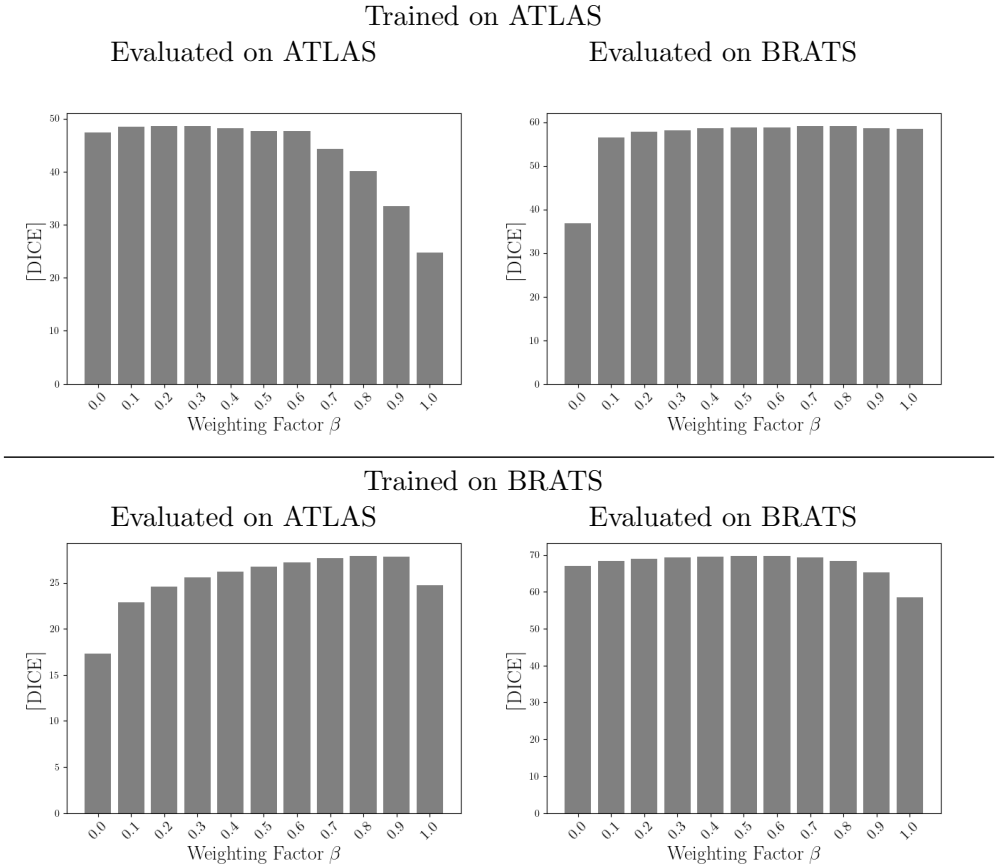


Figure 5: Analysis of the anomaly score weighting given
AS $= \beta \cdot x_{tilde} + (1 - \beta) \cdot \dot{\hat{y}}$, where x_{tilde} represents the anomaly map coming from the unsupervised reconstruction branch and $\dot{\hat{y}}$ represents the anomaly map coming from the supervised segmentation branch. The $\lceil$DICE$\rceil$ is plotted against different values of β.

List of Abbreviations

Abbreviations	Description
MRI	Magnetic Resonance Imaging
UAD	Unsupervised Anomaly Detection
GM	Generative Model
RF	Radio Frequency
FLAIR	Fluid Attenuated Inversion Recovery
MSE	Mean Squared Error
MAE	Mean Absolute Error
ANN	Artificial Neural Network
AN	Artificial Neuron
ReLU	Rectified Linear Unit
MLP	Multilayer Perceptron
CNN	Convolutional Neural Network
FC	Fully Connected
GAN	Generative Adversarial Network
AE	Autoencoder
VAE	Variational Autoencoder
ELBO	Evidence Lower Bound
DDPM	Denoising Diffusion Probabilistic Model
AD	Anomaly Detection
SVDD	Support Vector Data Description
SVM	Support Vector Machine
MN	MixedNormals
IXI	Information eXtraction from Images
BRATS	Brain Tumor Segmentation Challenge
MSLUB	Multiple Sclerosis data set from the University Hospital of Ljubljana
MS	Multiple Sclerosis
ATLAS	Anatomical Tracings of Lesions After Stroke
WMH	White Matter Hyperintensities
SSIM	Structural Similarity Index Measure
PSNR	Peak Signal-to-Noise Ratio
LPIPS	Learned Perceptual Image Patch Similarity
DICE	Sørensen-Dice Coefficient
AUPRC	Area Under the Precision-Recall Curve
FAE	Feature Autoencoder
RD	Reverse Distillation
EDC	Encoder-Decoder Consistency
SVAE	Sequential Variational Autoencoder
DAE	Denoising Autoencoder
RA	Reverse Anomaly

Abbreviations	Description
FPI	Foreign Patch Interpolation
Thresh	Thresholding
pDDPM	Patched Diffusion Model
cDDPM	Conditioned Diffusion Model
MHD	Mahalanobis Distance
SADM	Supervised Anomaly Detection with Diffusion Models
AS	Anomaly Score

Bibliography

[1] Börnert, P., Norris, D.G.: A half-century of innovation in technology-preparing mri for the 21st century. The British journal of radiology 93(1111), 20200113 (2020)

[2] Bruno, M.A., Walker, E.A., Abujudeh, H.H.: Understanding and confronting our mistakes: The epidemiology of error in radiology and strategies for error reduction. Radiographics : a review publication of the Radiological Society of North America, Inc 35(6), 1668–1676 (2015)

[3] Drew, T., Võ, M.L.H., Wolfe, J.M.: The invisible gorilla strikes again: sustained inattentional blindness in expert observers. Psychological science 24(9), 1848–1853 (2013)

[4] van Hespen, K.M., Zwanenburg, J.J.M., Dankbaar, J.W., Geerlings, M.I., Hendrikse, J., Kuijf, H.J.: An anomaly detection approach to identify chronic brain infarcts on mri. Scientific Reports 11(1), 7714 (2021)

[5] Bauer, S., Wiest, R., Nolte, L.P., Reyes, M.: A survey of mri-based medical image analysis for brain tumor studies. Physics in medicine and biology 58(13), R97–129 (2013)

[6] Porz, N., Bauer, S., Pica, A., Schucht, P., Beck, J., Verma, R.K., Slotboom, J., Reyes, M., Wiest, R.: Multi-modal glioblastoma segmentation: man versus machine. PloS one 9(5), e96873 (2014)

[7] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis 42, 60–88 (2017)

[8] Topol, E.J.: High-performance medicine: the convergence of human and artificial intelligence. Nature medicine 25(1), 44–56 (2019)

[9] Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.A., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, Ç., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz, J.A., Meier, R., Pereira, S., Precup, D., Price, S.J., Raviv, T.R., Reza, S.M.S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.C., Shotton, J., Silva, C.A., Sousa, N., Subbanna, N.K., Szekely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Unal, G., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Van Leemput, K.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging 34(10), 1993–2024 (2014)

[10] Maier, O., Menze, B.H., von der Gablentz, J., Hani, L., Heinrich, M.P., Liebrand, M., Winzeck, S., Basit, A., Bentley, P., Chen, L., Christiaens, D., Dutil, F., Egger, K., Feng, C., Glocker, B., Goetz, M., Haeck, T., Halme, H.L., Havaei, M., Iftekharuddin, K.M., Jodoin, P.M., Kamnitsas, K., Kellner, E., Korvenoja, A., Larochelle, H., Ledig, C., Lee, J.H., Maes, F., Mahmood, Q., Maier-Hein, K.H., McKinley, R., Muschelli, J., Pal, C., Pei, L., Rangarajan, J.R., Reza, S.M.S., Robben, D., Rueckert, D., Salli, E., Suetens, P., Wang, C.W., Wilms, M., Kirschke, J.S., Kr Amer, U.M., Münte, T.F., Schramm, P., Wiest, R., Handels, H., Reyes, M.: Isles 2015 - a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. Medical Image Analysis 35, 250–269 (2017)

[11] Carass, A., Roy, S., Jog, A., Cuzzocreo, J.L., Magrath, E., Gherman, A., Button, J., Nguyen, J., Prados, F., Sudre, C.H., Jorge Cardoso, M., Cawley, N., Ciccarelli, O., Wheeler-Kingshott, C.A.M., Ourselin, S., Catanese, L., Deshpande, H., Maurel, P., Commowick, O., Barillot, C., Tomas-Fernandez, X., Warfield, S.K., Vaidya, S., Chunduru, A., Muthuganapathy, R., Krishnamurthi, G., Jesson, A., Arbel, T., Maier, O., Handels, H., Iheme, L.O., Unay, D., Jain, S., Sima, D.M., Smeets, D., Ghafoorian, M., Platel, B., Birenbaum, A., Greenspan, H., Bazin, P.L., Calabresi, P.A., Crainiceanu, C.M., Ellingsen, L.M., Reich, D.S., Prince, J.L., Pham, D.L.: Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. NeuroImage 148, 77–102 (2017)

[12] Rauschecker, A.M., Rudie, J.D., Xie, L., Wang, J., Duong, M.T., Botzolakis, E.J., Kovalovich, A.M., Egan, J., Cook, T.C., Bryan, R.N., Nasrallah, I.M., Mohan, S., Gee, J.C.: Artificial intelligence system approaching neuroradiologist-level differential diagnosis accuracy at brain mri. Radiology 295(3), 626–637 (2020)

[13] Nagendran, M., Chen, Y., Lovejoy, C.A., Gordon, A.C., Komorowski, M., Harvey, H., Topol, E.J., Ioannidis, J.P., Collins, G.S., Maruthappu, M.: Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. bmj 368 (2020)

[14] Sogancioglu, E., Ginneken, B.v., Behrendt, F., Bengs, M., Schlaefer, A., Radu, M., Xu, D., Sheng, K., Scalzo, F., Marcus, E., Papa, S., Teuwen, J., Scholten, E.T., Schalekamp, S., Hendrix, N., Jacobs, C., Hendrix, W., Sánchez, C.I., Murphy, K.: Nodule detection and generation on chest x-rays: Node21 challenge. IEEE Transactions on Medical Imaging 43(8), 2839–2853 (2024)

[15] Ker, J., Wang, L., Rao, J., Lim, T.: Deep learning applications in medical image analysis. IEEE Access 6, 9375–9389 (2018)

[16] Aljuaid, A., Anwar, M.: Survey of supervised learning for medical image processing. SN Computer Science 3(4), 292 (2022)

[17] Li, M., Jiang, Y., Zhang, Y., Zhu, H.: Medical image analysis using deep learning algorithms. Frontiers in Public Health 11, 1273253 (2023)

[18] Decherchi, S., Pedrini, E., Mordenti, M., Cavalli, A., Sangiorgi, L.: Opportunities and challenges for machine learning in rare diseases. Frontiers in medicine 8, 747612 (2021)

[19] Vernooij, M.W., Ikram, M.A., Tanghe, H.L., Vincent, A.J.P.E., Hofman, A., Krestin, G.P., Niessen, W.J., Breteler, M.M.B., van der Lugt, A.: Incidental findings on brain mri in the general population. The New England journal of medicine 357(18), 1821–1828 (2007)

[20] Bengs, M., Behrendt, F., Krüger, J., Opfer, R., Schlaefer, A.: Three-dimensional deep learning with spatial erasing for unsupervised anomaly segmentation in brain mri. International journal of computer assisted radiology and surgery 16(9), 1413–1423 (2021)

[21] Behrendt, F., Bhattacharya, D., Krüger, J., Opfer, R., Schlaefer, A.: Patched diffusion models for unsupervised anomaly detection in brain mri. In: Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 227, 1019–1032 (2024)

[22] Behrendt, F., Bhattacharya, D., Mieling, R., Maack, L., Krüger, J., Opfer, R., Schlaefer, A.: Guided reconstruction with conditioned diffusion models for unsupervised anomaly detection in brain mris. Computers in Biology and Medicine 186, 109660 (2025)

[23] Behrendt, F., Bhattacharya, D., Maack, L., Krüger, J., Opfer, R., Mieling, R., Schlaefer, A.: Diffusion models with ensembled structure-based anomaly scoring for unsupervised anomaly detection. In: 2024 IEEE International Symposium on Biomedical Imaging. 1–4 (2024), © 2024 IEEE. Reprinted, with permission.

[24] Behrendt, F., Bhattacharya, D., Mieling, R., Maack, L., Krüger, J., Opfer, R., Schlaefer, A.: Leveraging the mahalanobis distance to enhance unsupervised brain mri anomaly detection. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. 394–404. Springer Nature Switzerland (2024), reproduced with permission from Springer Nature.

[25] Behrendt, F., Bhattacharya, D., Maack, L., Krüger, J., Opfer, R., Schlaefer, A.: Combining reconstruction-based unsupervised anomaly detection with supervised segmentation for brain mris. In: Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 250, 87–102 (2024)

[26] Weishaupt, D., Köchli, V.D., Marincek, B.: Wie funktioniert MRI?: eine Einführung in Physik und Funktionsweise der Magnetresonanzbildgebung. Springer Berlin Heidelberg (2009)

[27] Bakshi, R., Ariyaratana, S., Benedict, R.H., Jacobs, L.: Fluid-attenuated inversion recovery magnetic resonance imaging detects cortical and juxtacortical multiple sclerosis lesions. Archives of neurology 58(5), 742–748 (2001)

[28] Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., Prevedello, L.M., Rudie, J.D., Sako, C., Shinohara, R.T., Bergquist, T., Chai, R., Eddy, J., Elliott, J., Reade, W., Schaffter, T., Yu, T., Zheng, J., Moawad, A.W., Otavio Coelho, L., McDonnell, O., Miller, E., Moron, F.E., Oswood, M.C., Shih, R.Y., Siakallis, L., Bronstein, Y., Mason, J.R., Miller, A.F., Choudhary, G., Agarwal, A., Besada, C.H., Derakhshan,

J.J., Diogo, M.C., Do-Dai, D.D., Farage, L., Go, J.L., Hadi, M., Hill, V.B., Michael, I., Joyner, D., Lincoln, C., Lotan, E., Miyakoshi, A., Sanchez-Montano, M., Nath, J., Nguyen, X.V., Nicolas-Jilwan, M., Ortiz Jimenez, J., Ozturk, K., Petrovic, B.D., Shah, C., Shah, L.M., Sharma, M., Simsek, O., Singh, A.K., Soman, S., Statsevych, V., Weinberg, B.D., Young, R.J., Ikuta, I., Agarwal, A.K., Cambron, S.C., Silbergleit, R., Dusoi, A., Postma, A.A., Letourneau-Guillon, L., Guzman Perez-Carrillo, G.J., Saha, A., Soni, N., Zaharchuk, G., Zohrabian, V.M., Chen, Y., Cekic, M.M., Rahman, A., Small, J.E., Sethi, V., Davatzikos, C., Mongan, J., Hess, C., Cha, S., Villanueva-Meyer, J., Freymann, J.B., Kirby, J.S., Wiestler, B., Crivellaro, P., Colen, R.R., Kotrotsou, A., Marcus, D., Milchenko, M., Nazeri, A., Fathallah-Shaykh, H., Wiest, R., Jakab, A., Weber, M.A., Mahajan, A., Menze, B., Flanders, A.E., Bakas, S.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)

[29] Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. Adaptive Computation and Machine Learning, MIT Press, Cambridge, Massachusetts and London, England (2016)

[30] Pearson, K.: Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin philosophical magazine and journal of science 2(11), 559–572 (1901)

[31] Kramer, M.A.: Nonlinear principal component analysis using autoassociative neural networks. AIChE journal 37(2), 233–243 (1991)

[32] Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research 9(11) (2008)

[33] McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)

[34] Demuth, H.B., Beale, M.H., De Jess, O., Hagan, M.T.: Neural Network Design. Martin Hagan, Stillwater, OK, USA, 2nd edn. (2014)

[35] Fukushima, K.: Cognitron: A self-organizing multilayered neural network. Biological Cybernetics 20(3), 121–136 (1975)

[36] Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. 448–456 (2015)

[37] Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)

[38] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems 25 (2012)

[39] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[40] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1–9 (2015)

[41] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 770–778. Las Vegas, NV, USA (2016)

[42] Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2261–2269 (2017)

[43] Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. 6105–6114 (2019)

[44] Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollár, P.: Designing network design spaces. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 10428–10436 (2020)

[45] Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 11976–11986 (2022)

[46] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2015. 234–241 (2015)

[47] Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature 323(6088), 533–536 (1986)

[48] Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research 12(61), 2121–2159 (2011)

[49] Kingma, D., Welling, M.: Auto-encoding variational bayes. In: Bengio, Y., LeCun, Y. (eds.) 2nd International Conference on Learning Representations. ICLR 2014 (2014)

[50] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

[51] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. In: Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger (eds.) Proceedings of the 27th International Conference on Neural Information Processing Systems. 2672–2680 (2014)

[52] Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International conference on machine learning. 1530–1538 (2015)

[53] Van Den Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: International conference on machine learning. 1747–1756 (2016)

[54] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, 2256–2265. Lille, France (2015)

[55] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems 33, 6840–6851 (2020)

[56] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: 2017 IEEE International Conference on Computer Vision. 5908–5916 (2017)

[57] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 5967–5976 (2017)

[58] Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501 (2018)

[59] Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4681–4690 (2017)

[60] Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. In: The 25th biennial international conference on Information Processing in Medical Imaging (IPMI). 146–157. Springer International Publishing (2017)

[61] Chen, X., You, S., Tezcan, K.C., Konukoglu, E.: Unsupervised lesion detection via image restoration with a normative prior. Medical Image Analysis 64, 101713 (2020)

[62] Baur, C., Stefan Denner, Benedikt Wiestler, Nassir Navab, Shadi Albarqouni: Autoencoders for unsupervised anomaly segmentation in brain mr images: A comparative study. Medical Image Analysis 69, 101952 (2021)

[63] Luo, C.: Understanding diffusion models: A unified perspective. arXiv preprint arXiv:2208.11970 (2022)

[64] Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R.: Deconvolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2528–2535 (2010)

[65] Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning. 1096–1103 (2008)

[66] Bengio, Y., Yao, L., Alain, G., Vincent, P.: Generalized denoising auto-encoders as generative models. In: Proceedings of the 26th International Conference on Neural

Information Processing Systems. 899–907. Curran Associates Inc, Red Hook, NY, USA (2013)

[67] Alain, G., Bengio, Y.: What regularized auto-encoders learn from the data-generating distribution. Journal of Machine Learning Research 15(1), 3563–3593 (2014)

[68] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2016)

[69] Dilokthanakul, N., Mediano, P.A.M., Garnelo, M., Lee, M.C.H., Salimbeni, H., Arulkumaran, K., Shanahan, M.: Deep unsupervised clustering with gaussian mixture variational autoencoders. In: International Conference on Learning Representations (2017)

[70] Liu, L., Ren, Y., Lin, Z., Zhao, Z.: Pseudo numerical methods for diffusion models on manifolds. In: International Conference on Learning Representations (2022)

[71] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)

[72] Damudi, M.Z., Kini, A.S.: Single-step sampling approach for unsupervised anomaly detection of brain mri using denoising diffusion models. International Journal of Biomedical Imaging 2024(1), 2352602 (2024)

[73] Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM computing surveys 41(3), 1–58 (2009)

[74] Bhattacharyya, S., Jha, S., Tharakunnel, K., Westland, J.C.: Data mining for credit card fraud: A comparative study. Decision Support Systems 50(3), 602–613 (2011), on quantitative methods for detection of financial fraud

[75] Pourhabibi, T., Ong, K.L., Kam, B.H., Boo, Y.L.: Fraud detection: A systematic literature review of graph-based anomaly detection approaches. Decision Support Systems 133, 113303 (2020)

[76] Hilal, W., Gadsden, S.A., Yawney, J.: Financial fraud: a review of anomaly detection techniques and recent advances. Expert systems With applications 193, 116429 (2022)

[77] Tibshirani, R., Hastie, T.: Outlier sums for differential gene expression analysis. Biostatistics 8(1), 2–8 (2007)

[78] Kim, H., Gelenbe, E.: Anomaly detection in gene expression via stochastic models of gene regulatory networks. BMC genomics 10, 1–10 (2009)

[79] Czimmermann, T., Ciuti, G., Milazzo, M., Chiurazzi, M., Roccella, S., Oddo, C.M., Dario, P.: Visual-based defect detection and classification approaches for industrial applications—a survey. Sensors 20(5), 1459 (2020)

[80] Bergmann, P., Löwe, S., Fauser, M., Sattlegger, D., Steger, C.: Improving unsupervised defect segmentation by applying structural similarity to autoencoders. arXiv preprint arXiv:1807.02011 (2018)

[81] Bergmann, P., Fauser, M., Sattlegger, D., Steger, C.: Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4182–4191 (2020)

[82] Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9737–9746 (2022)

[83] Ruff, L., Kauffmann, J.R., Vandermeulen, R.A., Montavon, G., Samek, W., Kloft, M., Dietterich, T.G., Muller, K.R.: A unifying review of deep and shallow anomaly detection. Proceedings of the IEEE 109(5), 756–795 (2021)

[84] Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press (2001)

[85] Reynolds, D.: Gaussian Mixture Models, 827–832. Springer US, Boston, MA (2015)

[86] Rosenblatt, M.: Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics 27(3), 832–837 (1956)

[87] Parzen, E.: On estimation of a probability density function and mode. The Annals of Mathematical Statistics 33(3), 1065–1076 (1962)

[88] Reed, S., Oord, A., Kalchbrenner, N., Colmenarejo, S.G., Wang, Z., Chen, Y., Belov, D., Freitas, N.: Parallel multiscale autoregressive density estimation. In: International conference on machine learning. 2912–2921 (2017)

[89] Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the Support of a High-Dimensional Distribution. Neural Computation 13(7), 1443–1471 (2001)

[90] Tax, D.M.J., Duin, R.P.W.: Support vector data description. Machine Learning 54(1), 45–66 (2004)

[91] Moya, M.M., Hush, D.R.: Network constraints and multi-objective optimization for one-class classification. Neural Networks 9(3), 463–474 (1996)

[92] Menon, A.K., Williamson, R.C.: A loss framework for calibrated anomaly detection. In: Proceedings of the 32nd international conference on neural information processing systems. 1494–1504 (2018)

[93] Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, 4393–4402 (2018)

[94] Erfani, S.M., Rajasegarar, S., Karunasekera, S., Leckie, C.: High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. Pattern Recognition 58, 121–134 (2016)

[95] Xu, H., Chen, W., Zhao, N., Li, Z., Bu, J., Li, Z., Liu, Y., Zhao, Y., Pei, D., Feng, Y., Chen, J., Wang, Z., Qiao, H.: Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. In: Proceedings of the 2018 world wide web conference. 187–196 (2018)

[96] Martins, S., Barbara Caroline Benato, Bruna Ferreira Silva, Clarissa Lyn Yasuda, Alexandre Xavier Falcão: Modeling normal brain asymmetry in mr images applied to anomaly detection without segmentation and data annotation. In: Medical Imaging 2019: Computer-Aided Diagnosis. 71–80. SPIE (2019)

[97] Alaverdyan, Z., Jung, J., Bouet, R., Lartizien, C.: Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: Application to epilepsy lesion screening. Medical Image Analysis 60, 101618 (2020)

[98] Doorenbos, L., Sznitman, R., Márquez-Neila, P.: Ss3d: Unsupervised out-of-distribution detection and localization for medical volumes. In: Aubreville, M., Zimmerer, D., Heinrich, M. (eds.) Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis. 111–118. Springer International Publishing (2022)

[99] Lüth, C.T., Zimmerer, D., Koehler, G., Jaeger, P.F., Isenensee, F., Maier-Hein, K.H.: Contrastive representations for unsupervised anomaly detection and localization. In: BVM Workshop. 246–252. Springer (2023)

[100] Guo, J., Lu, S., Jia, L., Zhang, W., Li, H.: Encoder-decoder contrast for unsupervised anomaly detection in medical images. IEEE Transactions on Medical Imaging 43(3), 1102–1112 (2024)

[101] Meissen, F., Paetzold, J., Kaissis, G., Rueckert, D.: Unsupervised anomaly localization with structural feature-autoencoders. In: International MICCAI Brainlesion Workshop. 14–24. Springer (2022)

[102] Chen, X., Konukoglu, E.: Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. In: Medical Imaging with Deep Learning. Proceedings of Machine Learning Research (2018)

[103] Baur, C., Wiestler, B., Albarqouni, S., Navab, N.: Bayesian skip-autoencoders for unsupervised hyperintense anomaly detection in high resolution brain mri. In: 2020 IEEE 17th International Symposium on Biomedical Imaging. 1905–1909 (2020)

[104] Baur, C., Wiestler, B., Albarqouni, S., Navab, N.: Scale-space autoencoders for unsupervised anomaly segmentation in brain mri. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Lecture Notes in Computer Science, vol. 12264, 552–561. Springer International Publishing (2020)

[105] Aswani, K., Menaka, D.: A dual autoencoder and singular value decomposition based feature optimization for the segmentation of brain tumor from mri images. BMC Medical Imaging 21(1), 82 (2021)

[106] Baur, C., Wiestler, B., Muehlau, M., Zimmer, C., Navab, N., Albarqouni, S.: Modeling healthy anatomy with artificial intelligence for unsupervised anomaly detection in brain mri. Radiology: Artificial Intelligence 3(3), e190169 (2021)

[107] Uzunova, H., Schultz, S., Handels, H., Ehrhardt, J.: Unsupervised pathology detection in medical images using conditional variational autoencoders. International journal of computer assisted radiology and surgery 14(3), 451–461 (2019)

[108] Zimmerer, D., Isensee, F., Petersen, J., Kohl, S., Maier-Hein, K.: Unsupervised anomaly localization using variational auto-encoders. In: Shen, Dinggang and Liu, Tianming and Peters, Terry M. and Staib, Lawrence H. and Essert, Caroline and Zhou, Sean and Yap, Pew-Thian and Khan, Ali (ed.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. 289–297. Springer International Publishing (2019)

[109] Zimmerer, D., Kohl, S., Petersen, J., Isensee, F., Maier-Hein, K.: Context-encoding variational autoencoder for unsupervised anomaly detection. In: Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 102 (2019)

[110] Akrami, H., Joshi, A.A., Li, J., Aydore, S., Leahy, R.M.: Brain lesion detection using a robust variational autoencoder and transfer learning. IEEE 17th International Symposium on Biomedical Imaging 786–790 (2020)

[111] Baur, C., Wiestler, B., Albarqouni, S., Navab, N.: Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Image Processing, Computer Vision, Pattern Recognition, and Graphics, vol. 11383, 161–169. Springer International Publishing (2019)

[112] Baur, C., Graf, R., Wiestler, B., Albarqouni, S., Navab, N.: Steganomaly: Inhibiting cyclegan steganography for unsupervised anomaly detection in brain mri. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Lecture Notes in Computer Science, vol. 12262, 718–727. Springer International Publishing (2020)

[113] Nguyen, B., Feldman, A., Bethapudi, S., Jennings, A., Willcocks, C.G.: Unsupervised region-based anomaly detection in brain mri with adversarial image inpainting. In: 2021 IEEE 18th International Symposium on Biomedical Imaging. 1127–1131 (2021)

[114] Han, C., Rundo, L., Murao, K., Noguchi, T., Shimahara, Y., Milacski, Z.Á., Koshino, S., Sala, E., Nakayama, H., Satoh, S.: Madgan: unsupervised medical anomaly detection gan using multiple adjacent brain mri slice reconstruction. BMC bioinformatics 22, 31 (2021)

[115] Kascenas, A., Pugeault, N., O'Neil, A.Q.: Denoising autoencoders for unsupervised anomaly detection in brain mri. In: Konukoglu, E., Menze, B., Venkataraman, A., Baumgartner, C., Dou, Q., Albarqouni, S. (eds.) Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 172, 653–664 (2022)

[116] Wyatt, J., Leach, A., Schmon, S.M., Willcocks, C.G.: Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 650–656 (2022)

[117] Pinaya, W.H.L., Graham, M.S., Gray, R., da Costa, P.F., Tudosiu, P.D., Wright, P., Mah, Y.H., MacKinnon, A.D., Teo, J.T., Jager, R., Werring, D., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J.: Fast unsupervised brain anomaly detection and segmentation with diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. 705–714 (2022)

[118] Zimmerer, D., Full, P.M., Isensee, F., Jager, P., Adler, T., Petersen, J., Kohler, G., Ross, T., Reinke, A., Kascenas, A., Jensen, B.S., O'Neil, A.Q., Tan, J., Hou, B., Batten, J., Qiu, H., Kainz, B., Shvetsova, N., Fedulova, I., Dylov, D.V., Yu, B., Zhai, J., Hu, J., Si, R., Zhou, S., Wang, S., Li, X., Chen, X., Zhao, Y., Marimont, S.N., Tarroni, G., Saase, V., Maier-Hein, L., Maier-Hein, K.: Mood 2020: A public benchmark for out-of-distribution detection and localization on medical images. IEEE transactions on medical imaging 41(10), 2728–2738 (2022)

[119] Tan, J., Hou, B., Day, T., Simpson, J., Rueckert, D., Kainz, B.: Detecting outliers with poisson image interpolation. 581–591. Springer, Cham (2021)

[120] Tan, J., Hou, B., Batten, J., Qiu, H., Kainz, B.: Detecting outliers with foreign patch interpolation. Machine Learning for Biomedical Imaging 1, 1–27 (2022)

[121] Baugh, M., Tan, J., Müller, J.P., Dombrowski, M., Batten, J., Kainz, B.: Many tasks make light work: Learning to localise medical anomalies from multiple synthetic tasks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. 162–172. Springer (2023)

[122] Biomedical Image Analysis Group: Ixi dataset – brain development

[123] Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data 4(1), 1–13 (2017)

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

[125] Liew, S.L., Lo, B.P., Donnelly, M.R., Zavaliangos-Petropulu, A., Jeong, J.N., Barisano, G., Hutton, A., Simon, J.P., Juliano, J.M., Suri, A., Wang, Z., Abdullah, A., Kim, J., Ard, T., Banaj, N., Borich, M.R., Boyd, L.A., Brodtmann, A., Buetefisch, C.M., Cao, L., Cassidy, J.M., Ciullo, V., Conforto, A.B., Cramer, S.C., Dacosta-Aguayo, R., de la Rosa, E., Domin, M., Dula, A.N., Feng, W., Franco, A.R., Geranmayeh, F., Gramfort, A., Gregory, C.M., Hanlon, C.A., Hordacre, B.G., Kautz, S.A., Khlif, M.S., Kim, H., Kirschke, J.S., Liu, J., Lotze, M., MacIntosh, B.J., Mataró, M., Mohamed, F.B., Nordvik, J.E., Park, G., Pienta, A., Piras, F., Redman, S.M., Revill, K.P., Reyes, M., Robertson, A.D., Seo, N.J., Soekadar, S.R.,

Spalletta, G., Sweet, A., Telenczuk, M., Thielman, G., Westlye, L.T., Winstein, C.J., Wittenberg, G.F., Wong, K.A., Yu, C.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Scientific data 9(1), 320 (2022)

[126] Kuijf, H.J., Biesbroek, J.M., De Bresser, J., Heinen, R., Andermatt, S., Bento, M., Berseth, M., Belyaev, M., Cardoso, M.J., Casamitjana, A., Collins, D.L., Dadar, M., Georgiou, A., Ghafoorian, M., Jin, D., Khademi, A., Knight, J., Li, H., Lladó, X., Luna, M., Mahmood, Q., McKinley, R., Mehrtash, A., Ourselin, S., Park, B.Y., Park, H., Park, S.H., Pezold, S., Puybareau, E., Rittner, L., Sudre, C.H., Valverde, S., Vilaplana, V., Wiest, R., Xu, Y., Xu, Z., Zeng, G., Zhang, J., Zheng, G., Chen, C., van der Flier, W., Barkhof, F., Viergever, M.A., Biessels, G.J.: Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. IEEE transactions on medical imaging 38(11), 2556–2568 (2019)

[127] Rohlfing, T., Zahr, N.M., Sullivan, E.V., Pfefferbaum, A.: The sri24 multichannel atlas of normal adult human brain structure. Human brain mapping 31(5), 798–819 (2010)

[128] Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.P., Heiland, S., Wick, W., Bendszus, M., Maier-Hein, K.H., Kickingereder, P.: Automated brain extraction of multisequence mri using artificial neural networks. Human brain mapping 40(17), 4952–4964 (2019)

[129] Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4itk: improved n3 bias correction. IEEE transactions on medical imaging 29(6), 1310–1320 (2010)

[130] Tustison, N.J., Cook, P.A., Holbrook, A.J., Johnson, H.J., Muschelli, J., Devenyi, G.A., Duda, J.T., Das, S.R., Cullen, N.C., Gillen, D.L., Yassa, M.A., Stone, J.R., Gee, J.C., Avants, B.B.: The antsx ecosystem for quantitative biological and medical imaging. Scientific Reports 11(1), 9068 (2021)

[131] Pérez-García, F., Sparks, R., Ourselin, S.: Torchio: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Computer Methods and Programs in Biomedicine 208, 106236 (2021)

[132] Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13(4), 600–612 (2004)

[133] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 586–595 (2018)

[134] Dice, L.R.: Measures of the amount of ecologic association between species. Ecology 26, 297–302 (1945)

[135] Sørensen, T., Sørensen, T., Biering-Sørensen, T., Sørensen, T., Sorensen, J.T.: A method of establishing group of equal amplitude in plant sociobiology based on similarity of species content and its application to analyses of the vegetation on danish commons. Kongelige Danske Videnskabernes Selskab 5, 1–34 (1948)

[136] Raschka, S.: Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. The Journal of Open Source Software 3(24) (2018)

[137] Behrendt, F., Bengs, M., Bhattacharya, D., Krüger, J., Opfer, R., Schlaefer, A.: Capturing inter-slice dependencies of 3d brain mri-scans for unsupervised anomaly detection. In: Medical Imaging with Deep Learning. Proceedings of Machine Learning Research (2022)

[138] Bercea, C., Benedikt Wiestler, Daniel Rueckert, Julia A Schnabel: Generalizing unsupervised anomaly detection: Towards unbiased pathology screening. In: Medical Imaging with Deep Learning. Proceedings of Machine Learning Research (2023)

[139] Daniel, T., Tamar, A.: Soft-introvae: Analyzing and improving the introspective variational autoencoder. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4391–4400 (2021)

[140] Bercea, C.I., Wiestler, B., Rueckert, D., Schnabel, J.A.: Reversing the abnormal: Pseudo-healthy generative networks for anomaly detection. In: Medical Image Computing and Computer Assisted Intervention – MICCAI23. vol. 14224, 293–303 (2023)

[141] Zavrtanik, V., Kristan, M., Skočaj, D.: Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 8330–8339 (2021)

[142] Meissen, F., Kaissis, G., Rueckert, D.: Challenging current semi-supervised anomaly segmentation methods for brain mri. In: Crimi, Alessandro and Bakas, Spyridon (ed.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. 63–74. Springer International Publishing (2022)

[143] Bredell, G., Flouris, K., Chaitanya, K., Erdil, E., Konukoglu, E.: Explicitly minimizing the blur error of variational autoencoders. In: International Conference on Learning Representations (2022)

[144] Lagogiannis, I., Meissen, F., Kaissis, G., Rueckert, D.: Unsupervised pathology detection: A deep dive into the state of the art. IEEE transactions on medical imaging PP (2023)

[145] DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)

[146] Lambert, B., Louis, M., Doyle, S., Forbes, F., Dojat, M., Tucholka, A.: Leveraging 3d information in unsupervised brain mri segmentation. In: 2021 IEEE 18th International Symposium on Biomedical Imaging. 187–190 (2021)

[147] Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: AAAI. vol. 32 (2018)

[148] Tian, K., Jiang, Y., qishuai diao, Lin, C., Wang, L., Yuan, Z.: Designing BERT for convolutional networks: Sparse and hierarchical masked modeling. In: International Conference on Learning Representations (2023)

[149] Meissen, F., Wiestler, B., Kaissis, G., Rueckert, D.: On the pitfalls of using the residual error as anomaly score. Proceedings of Machine Learning Research, vol. 172, 1—-15 (2022)

[150] Sato, K., Hama, K., Matsubara, T., Uehara, K.: Predictable uncertainty-aware unsupervised deep anomaly segmentation. In: 2019 International Joint Conference on Neural Networks. 1–7. IEEE, Piscataway, NJ (2019)

[151] Mao, Y., Xue, F.F., Wang, R., Zhang, J., Zheng, W.S., Liu, H.: Abnormality detection in chest x-ray images using uncertainty prediction autoencoders. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2020. 529–538. Springer (2020)

[152] Mahalanobis, P.: On the generalised distance in statistics. In: Proceedings of the National Institute of Science of India. vol. 12, 49–55 (1936)

[153] Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., Prastawa, M., Alberts, E., Lipkova, J., Freymann, J., Kirby, J., Bilello, M., Fathallah-Shaykh, H., Wiest, R., Kirschke, J., Wiestler, B., Colen, R., Kotrotsou, A., Lamontagne, P., Marcus, D., Milchenko, M., Nazeri, A., Weber, M.A., Mahajan, A., Baid, U., Gerstner, E., Kwon, D., Acharya, G., Agarwal, M., Alam, M., Albiol, A., Albiol, A., Albiol, F.J., Alex, V., Allinson, N., Amorim, P.H., Amrutkar, A., Anand, G., Andermatt, S., Arbel, T., Arbelaez, P., Avery, A., Azmat, M., Pranjal, B., Bai, W., Banerjee, S., Barth, B., Batchelder, T., Batmanghelich, K., Battistella, E., Beers, A., Belyaev, M., Bendszus, M., Benson, E., Bernal, J., Bharath, H.N., Biros, G., Bisdas, S., Brown, J., Cabezas, M., Cao, S., Cardoso, J.M., Carver, E.N., Casamitjana, A., Castillo, L.S., Catà, M., Cattin, P., Cerigues, A., Chagas, V.S., Chandra, S., Chang, Y.J., Chang, S., Chang, K., Chazalon, J., Chen, S., Chen, W., Chen, J.W., Chen, Z., Cheng, K., Choudhury, A.R., Chylla, R., Clérigues, A., Colleman, S., Colmeiro, R.G.R., Combalia, M., Costa, A., Cui, X., Dai, Z., Dai, L., Daza, L.A., Deutsch, E., Ding, C., Dong, C., Dong, S., Dudzik, W., Eaton-Rosen, Z., Egan, G., Escudero, G., Estienne, T., Everson, R., Fabrizio, J., Fan, Y., Fang, L., Feng, X., Ferrante, E., Fidon, L., Fischer, M., French, A.P., Fridman, N., Fu, H., Fuentes, D., Gao, Y., Gates, E., Gering, D., Gholami, A., Gierke, W., Glocker, B., Gong, M., González-Villá, S., Grosges, T., Guan, Y., Guo, S., Gupta, S., Han, W.S., Han, I.S., Harmuth, K., He, H., Hernández-Sabaté, A., Herrmann, E., Himthani, N., Hsu, W., Hsu, C., Hu, X., Hu, X., Hu, Y., Hu, Y., Hua, R., Huang, T.Y., Huang, W., Huffel, S.V., Huo, Q., Vivek, H., Iftekharuddin, K.M., Isensee, F., Islam, M., Jackson, A.S., Jambawalikar, S.R., Jesson, A., Jian, W., Jin, P., Jose, V.J.M., Jungo, A., Kainz, B., Kamnitsas, K., Kao, P.Y., Karnawat, A., Kellermeier, T., Kermi, A., Keutzer, K., Khadir, M.T., Khened, M., Kickingereder, P., Kim, G., King, N., Knapp, H., Knecht, U., Kohli, L., Kong, D., Kong, X., Koppers, S., Kori, A., Krishnamurthi, G., Krivov, E., Kumar, P., Kushibar, K., Lachinov, D., Lambrou, T., Lee, J., Lee, C., Lee, Y., Lee, M., Lefkovits, S., Lefkovits, L., Levitt, J., Li, T., Li, H., Li, W., Li, H., Li, X., Li, Y., Li, H., Li, Z., Li, X., Li, Z., Li, X., Li, W., Lin, Z.S., Lin, F.,

Lio, P., Liu, C., Liu, B., Liu, X., Liu, M., Liu, J., Liu, L., Llado, X., Lopez, M.M., Lorenzo, P.R., Lu, Z., Luo, L., Luo, Z., Ma, J., Ma, K., Mackie, T., Madabushi, A., Mahmoudi, I., Maier-Hein, K.H., Maji, P., Mammen, C., Mang, A., Manjunath, B., Marcinkiewicz, M., McDonagh, S., McKenna, S., McKinley, R., Mehl, M., Mehta, S., Mehta, R., Meier, R., Meinel, C., Merhof, D., Meyer, C., Miller, R., Mitra, S., Moiyadi, A., Molina-Garcia, D., Monteiro, M.A., Mrukwa, G., Myronenko, A., Nalepa, J., Ngo, T., Nie, D., Ning, H., Niu, C., Nuechterlein, N.K., Oermann, E., Oliveira, A., Oliveira, D.D., Oliver, A., Osman, A.F., Ou, Y.N., Ourselin, S., Paragios, N., Park, M.S., Paschke, B., Pauloski, J.G., Pawar, K., Pawlowski, N., Pei, L., Peng, S., Pereira, S.M., Perez-Beteta, J., Perez-Garcia, V.M., Pezold, S., Pham, B., Phophalia, A., Piella, G., Pillai, G., Piraud, M., Pisov, M., Popli, A., Pound, M.P., Pourreza, R., Prasanna, P., Prkovska, V., Pridmore, T.P., Puch, S., Puybareau, É., Qian, B., Qiao, X., Rajchl, M., Rane, S., Rebsamen, M., Ren, H., Ren, X., Revanuru, K., Rezaei, M., Rippel, O., Rivera, L.C., Robert, C., Rosen, B., Rueckert, D., Safwan, M., Salem, M., Salvi, J., Sanchez, I., Sánchez, I., Santos, H.M., Sartor, E., Schellingerhout, D., Scheufele, K., Scott, M.R., Scussel, A.A., Sedlar, S., Serrano-Rubio, J.P., Shah, N.J., Shah, N., Shaikh, M., Shankar, B.U., Shboul, Z., Shen, H., Shen, D., Shen, L., Shen, H., Shenoy, V., Shi, F., Shin, H.E., Shu, H., Sima, D., Sinclair, M., Smedby, O., Snyder, J.M., Soltaninejad, M., Song, G., Soni, M., Stawiaski, J., Subramanian, S., Sun, L., Sun, R., Sun, J., Sun, K., Sun, Y., Sun, G., Sun, S., Suter, Y.R., Szilagyi, L., Talbar, S., Tao, D., Tao, D., Teng, Z., Thakur, S., Thakur, M.H., Tharakan, S., Tiwari, P., Tochon, G., Tran, T., Tsai, Y.M., Tseng, K.L., Tuan, T.A., Turlapov, V., Tustison, N., Vakalopoulou, M., Valverde, S., Vanguri, R., Vasiliev, E., Ventura, J., Vera, L., Vercauteren, T., Verrastro, C., Vidyaratne, L., Vilaplana, V., Vivekanandan, A., Wang, G., Wang, Q., Wang, C.J., Wang, W., Wang, D., Wang, R., Wang, Y., Wang, C., Wang, G., Wen, N., Wen, X., Weninger, L., Wick, W., Wu, S., Wu, Q., Wu, Y., Xia, Y., Xu, Y., Xu, X., Xu, P., Yang, T.L., Yang, X., Yang, H.Y., Yang, J., Yang, H., Yang, G., Yao, H., Ye, X., Yin, C., Young-Moxon, B., Yu, J., Yue, X., Zhang, S., Zhang, A., Zhang, K., Zhang, X., Zhang, L., Zhang, X., Zhang, Y., Zhang, L., Zhang, J., Zhang, X., Zhang, T., Zhao, S., Zhao, Y., Zhao, X., Zhao, L., Zheng, Y., Zhong, L., Zhou, C., Zhou, X., Zhou, F., Zhu, H., Zhu, J., Zhuge, Y., Zong, W., Kalpathy-Cramer, J., Farahani, K., Davatzikos, C., Leemput, K.V., Menze, B.: 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)

[154] McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F.J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C.J., King, D., Ledsam, J.R., Melnick, D., Mostofi, H., Peng, L., Reicher, J.J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K.C., De Fauw, J., Shetty, S.: International evaluation of an ai system for breast cancer screening. Nature 577(7788), 89–94 (2020)

[155] Rodríguez-Ruiz, A., Krupinski, E., Mordang, J.J., Schilling, K., Heywang-Köbrunner, S.H., Sechopoulos, I., Mann, R.M.: Detection of breast cancer with

mammography: effect of an artificial intelligence support system. Radiology 290(2), 305–314 (2019)

[156] Ardila, D., Kiraly, A.P., Bharadwaj, S., Choi, B., Reicher, J.J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., Naidich, D.P., Shetty, S.: End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine 25(6), 954–961 (2019)

[157] Bercea, C.I., Wiestler, B., Rueckert, D., Albarqouni, S.: Federated disentangled representation learning for unsupervised brain anomaly detection. Nature Machine Intelligence 4(8), 685–695 (2022)

[158] Bengs, M., Behrendt, F., Max-Heinrich Laves, Krüger, J., Opfer, R., Schlaefer, A.: Unsupervised anomaly detection in 3d brain mri using deep learning with multi-task brain age prediction. In: Karen Drukker, Khan M. Iftekharuddin (eds.) Medical Imaging 2022: Computer-Aided Diagnosis. vol. 12033, 1203314. SPIE (2022)

[159] Cai, Y., Chen, H., Yang, X., Zhou, Y., Cheng, K.T.: Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images. Medical Image Analysis 86, 102794 (2023)

[160] Marimont, S.N., Tarroni, G.: Anomaly detection through latent space restoration using vector quantized variational autoencoders. In: 2021 IEEE 18th International Symposium on Biomedical Imaging. 1764–1767 (2021)

[161] Pinaya, W.H.L., Tudosiu, P.D., Gray, R., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J.: Unsupervised brain anomaly detection and segmentation with transformers. In: Heinrich, M., Dou, Q., de Bruijne, M., Lellmann, J., Schläfer, A., Ernst, F. (eds.) Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 143, 596–617 (2021)

[162] Pinaya, W.H.L., Tudosiu, P.D., Gray, R., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J.: Unsupervised brain imaging 3d anomaly detection and segmentation with transformers. Medical Image Analysis 79, 102475 (2022)

[163] Kascenas, A., Sanchez, P., Schrempf, P., Wang, C., Clackett, W., Mikhael, S.S., Voisey, J.P., Goatman, K., Weir, A., Pugeault, N., Tsaftaris, S.A., O'Neil, A.Q.: The role of noise in denoising models for anomaly detection in medical images. Medical Image Analysis 90, 102963 (2023)

[164] Iqbal, H., Khalid, U., Chen, C., Hua, J.: Unsupervised anomaly detection in medical images using masked diffusion model. In: Cao, X., Xu, X., Rekik, I., Cui, Z., Ouyang, X. (eds.) Machine Learning in Medical Imaging, Lecture Notes in Computer Science, vol. 14348, 372–381. Springer Nature Switzerland, Cham (2023)

[165] Bercea, C., Michael Neumayr, Daniel Rueckert, Julia A Schnabel: Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models. In: ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH) (2023)

[166] Graham, M.S., Pinaya, W.H., Tudosiu, P.D., Nachev, P., Ourselin, S., Cardoso, J.: Denoising diffusion models for out-of-distribution detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2948–2957 (2023)

[167] Bercea, C.I., Wiestler, B., Rueckert, D., Schnabel, J.A.: Diffusion models with implicit guidance for medical anomaly detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. 211–220. Springer (2024)

[168] Wolleb, J., Bieder, F., Friedrich, P., Zhang, P., Durrer, A., Cattin, P.C.: Binary noise for binary tasks: Masked bernoulli diffusion for unsupervised anomaly detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. 135–145. Springer (2024)

[169] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision. 618–626 (2017)

[170] Silva-Rodríguez, J., Naranjo, V., Dolz, J.: Constrained unsupervised anomaly segmentation. Medical Image Analysis 80, 102526 (2022)

[171] Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems 31 (2018)

[172] Vasiliuk, A., Frolova, D., Belyaev, M., Shirokikh, B.: Limitations of out-of-distribution detection in 3d medical image segmentation. Journal of Imaging 9(9) (2023)

[173] Saase, V., Wenz, H., Ganslandt, T., Groden, C., Maros, M.E.: Simple statistical methods for unsupervised brain anomaly detection on mri are competitive to deep learning methods. arXiv preprint arXiv:2011.12735 (2020)

[174] Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C.: Diffusion models for medical anomaly detection. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. vol. 13438, 35–45. Springer Nature Switzerland (2022)

[175] Sanchez, P., Kascenas, A., Liu, X., O'Neil, A.Q., Tsaftaris, S.A.: What is healthy? generative counterfactual diffusion for lesion localization. In: MICCAI Workshop on Deep Generative Models. 34–44. Springer (2022)

[176] Fontanella, A., Mair, G., Wardlaw, J., Trucco, E., Storkey, A.: Diffusion models for counterfactual generation and anomaly detection in brain images. IEEE Transactions on Medical Imaging (2024)

[177] Marimont, S.N., Baugh, M., Siomos, V., Tzelepis, C., Kainz, B., Tarroni, G.: Disyre: Diffusion-inspired synthetic restoration for unsupervised anomaly detection. In: 2024 IEEE International Symposium on Biomedical Imaging. 1–5 (2024)

[178] Naval Marimont, S., Siomos, V., Baugh, M., Tzelepis, C., Kainz, B., Tarroni, G.: Ensembled cold-diffusion restorations for unsupervised anomaly detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. 243–253. Springer (2024)

[179] Bansal, A., Borgnia, E., Chu, H.M., Li, J., Kazemi, H., Huang, F., Goldblum, M., Geiping, J., Goldstein, T.: Cold diffusion: Inverting arbitrary image transforms without noise. Advances in Neural Information Processing Systems 36 (2024)

[180] Baugh, M., Reynaud, H., Marimont, S.N., Cechnicka, S., Müller, J.P., Tarroni, G., Kainz, B.: Image-conditioned diffusion models for medical anomaly detection. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. 117–127. Springer (2024)

[181] Behrendt, F., Bengs, M., Rogge, F., Kruger, J., Opfer, R., Schlaefer, A.: Unsupervised anomaly detection in 3d brain mri using deep learning with impured training data. In: 2022 IEEE 19th International Symposium on Biomedical Imaging. 1–4 (2022)

[182] Siddiquee, M.M.R., Shah, J., Wu, T., Chong, C., Schwedt, T.J., Dumkrieger, G., Nikolova, S., Li, B.: Brainomaly: Unsupervised neurologic disease detection utilizing unannotated t1-weighted brain mr images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 7573–7582 (2024)

[183] Chen, X., Pawlowski, N., Rajchl, M., Glocker, B., Konukoglu, E.: Deep generative models in the real-world: An open challenge from medical imaging. arXiv preprint 1806.05452v1 (2018)

[184] Akrami, H., Joshi, A., Aydore, S., Leahy, R.: Deep quantile regression for uncertainty estimation in unsupervised and supervised lesion detection. Machine Learning for Biomedical Imaging 1(IPMI 2021), 1–23 (2022)

[185] Bercea, C.I., Wiestler, B., Rueckert, D., Schnabel, J.A.: Evaluating normative representation learning in generative ai for robust anomaly detection in brain imaging. Nature Communications 16(1), 1624 (2025)

[186] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. 10684–10695 (2022)

[187] Wolleb, J., Bieder, F., Friedrich, P., Zhang, P., Durrer, A., Cattin, P.C.: Binary noise for binary tasks: Masked bernoulli diffusion for unsupervised anomaly detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. 135–145. Springer (2024)

[188] Friedrich, P., Wolleb, J., Bieder, F., Durrer, A., Cattin, P.C.: Wdm: 3d wavelet diffusion models for high-resolution medical image synthesis. In: MICCAI Workshop on Deep Generative Models. 11–21. Springer (2024)

[189] Jagodzinski, A., Johansen, C., Koch-Gromus, U., Aarabi, G., Adam, G., Anders, S., Augustin, M., der Kellen, R.B., Beikler, T., Behrendt, C.A., Betz, C.S., Bokemeyer, C., Borof, K., Briken, P., Busch, C.J., Büchel, C., Brassen, S., Debus, E.S., Eggers, L., Fiehler, J., Gallinat, J., Gellißen, S., Gerloff, C., Girdauskas, E., Gosau, M.,

Graefen, M., Härter, M., Harth, V., Heidemann, C., Heydecke, G., Huber, T.B., Hussein, Y., Kampf, M.O., von dem Knesebeck, O., Konnopka, A., König, H.H., Kromer, R., Kubisch, C., Kühn, S., Loges, S., Löwe, B., Lund, G., Meyer, C., Nagel, L., Nienhaus, A., Pantel, K., Petersen, E., Püschel, K., Reichenspurner, H., Sauter, G., Scherer, M., Scherschel, K., Schiffner, U., Schnabel, R.B., Schulz, H., Smeets, R., Sokalskis, V., Spitzer, M.S., Terschüren, C., Thederan, I., Thoma, T., Thomalla, G., Waschki, B., Wegscheider, K., Wenzel, J.P., Wiese, S., Zyriax, B.C., Zeller, T., Blankenberg, S.: Rationale and design of the hamburg city health study. European Journal of Epidemiology 35(2), 169–181 (2020)

[190] Ma, Z., Reich, D.S., Dembling, S., Duyn, J.H., Koretsky, A.P.: Outlier detection in multimodal mri identifies rare individual phenotypes among more than 15,000 brains. Human brain mapping 43(5), 1766–1782 (2022)