

Deep Unsupervised Anomaly Detection in Brain Imaging: Large-Scale Benchmarking and Bias Analysis

Alexander Frotscher^a, Christian F. Baumgartner^{b,e}, Thomas Wolfers^{a,c,d}

^aDepartment of Psychiatry and Psychotherapy, University Hospital Tübingen, Tübingen, Germany

^bCluster of Excellence – Machine Learning for Science, University of Tübingen, Baden-Württemberg, Germany

^cDepartment of Psychology, Friedrich Schiller University of Jena, Germany

^dGerman Center for Mental Health (DZPG), partner site Halle/Jena/Magdeburg, Germany

^eUniversity of Lucerne, Lucerne, Switzerland

December 2, 2025

Abstract

Deep unsupervised anomaly detection in brain magnetic resonance imaging offers a promising route to identify pathological deviations without requiring lesion-specific annotations. Yet, fragmented evaluations, heterogeneous datasets, and inconsistent metrics have hindered progress toward clinical translation. Here, we present a large-scale, multi-center benchmark of deep unsupervised anomaly detection for brain imaging. The training cohort comprised 2,976 T1 and 2,972 T2-weighted scans ($\approx 461,000$ slices) from healthy individuals across six scanners, with ages ranging from 6 to 89 years. Validation used 92 scans to tune hyperparameters and estimate unbiased thresholds. Testing encompassed 2,221 T1w and 1,262 T2w scans spanning healthy datasets and diverse clinical cohorts. Across all algorithms, the Dice-based segmentation performance varied between ≈ 0.03 and ≈ 0.65 , indicating substantial variability and underscoring that no single method achieved consistent superiority across lesion types or modalities for any task. To assess robustness, we systematically evaluated the impact of different scanners, lesion types and sizes, as well as demographics (age, sex). *Reconstruction-based* methods, particularly diffusion-inspired approaches, achieved the strongest lesion segmentation performance, while *feature-based* meth-

ods showed greater robustness under distributional shifts. However, systematic biases, such as scanner-related effects, were observed for the majority of algorithms, including that small and low-contrast lesions were missed more often, and that false positives varied with age and sex. Increasing healthy training data yields only modest gains, underscoring that current unsupervised anomaly detection frameworks are limited algorithmically rather than by data availability. Our benchmark establishes a transparent foundation for future research and highlights priorities for clinical translation, including image native pretraining, principled deviation measures, fairness-aware modeling, and robust domain adaptation.

1 Introduction

The human brain is the most complex organ in the body, with billions of neurons forming intricate networks that support cognition, behavior, and perception [18;24;39](#). Neuroimaging provides insights into this complexity but necessarily compresses rich biology into high-dimensional, often noisy measurements. Diseases of neurodegeneration, mental disorders, stroke, and tumors are characterized by substantial heterogeneity at the brain level [17](#). This heterogeneity arises not only from the underlying pathology itself but is further amplified by demo-

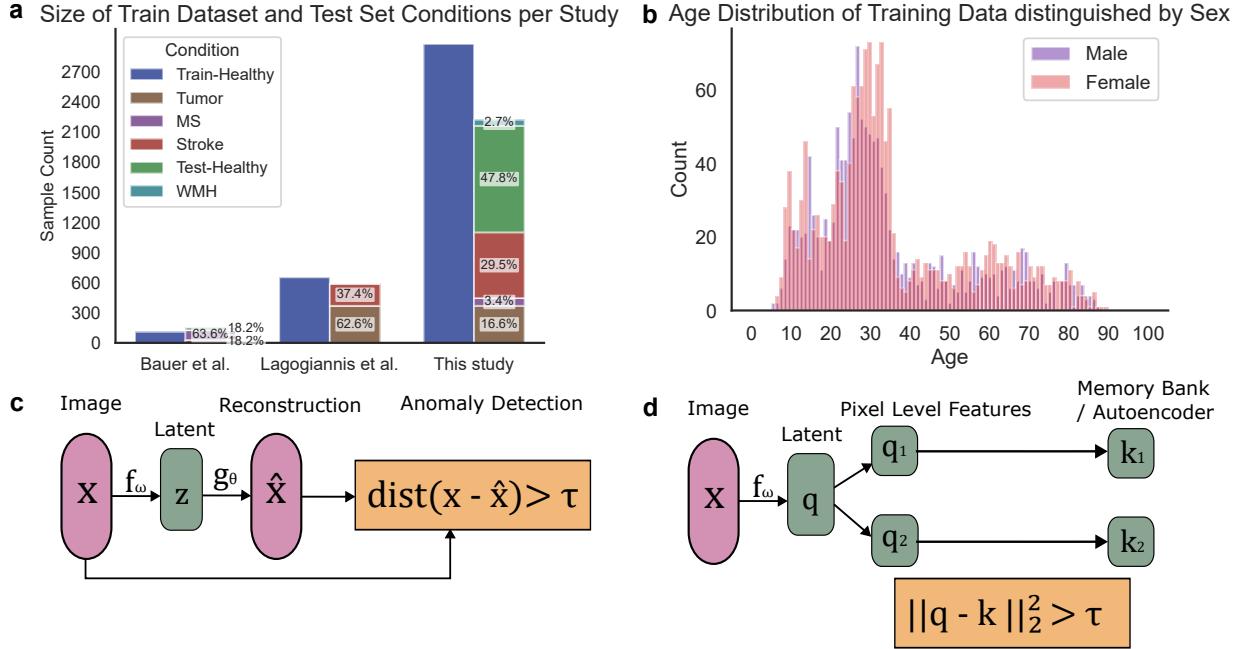


Figure 1: **Large-scale benchmarking of interdisciplinary state-of-the-art deep unsupervised anomaly detection.** **a)** Number of volumes in the training and test datasets, split by condition, compared to previous benchmarks. **b)** Age distribution of the training data. **c)** Example of a *reconstruction-based* approach. An encoder and decoder are trained exclusively on healthy data, and during testing anomalies are identified by thresholding the reconstruction error with τ . **d)** Example of a *feature-based* approach. A feature map is created by passing the image through a neural network. These features are either (i) reconstructed using an autoencoder, analogous to panel c, or (ii) compared directly to features from a memory bank of healthy images. Anomalies are flagged when the Euclidean distance to healthy features exceeds τ .

graphic factors such as age and sex, genetic predispositions, environmental influences, and lifestyle factors. Moreover, technical variability stemming from magnetic resonance imaging (MRI) acquisition protocols, scanner hardware, and image quality^{38;76} adds another layer of complexity, making robust characterization and analysis particularly challenging. In clinical practice, diagnosis of brain lesions and treatment planning generally rely on human evaluation of MRI, including lesion identification^{5;73}. Yet, even for experienced clinicians, the heterogeneity of disease-related changes makes manual assessment time-consuming and error-prone⁸². To reduce misdiagnosis, streamline planning, enable longitudinal monitoring, and alleviate workload, auto-

mated lesion detection and decision-support systems have emerged^{15;35}. State-of-the-art systems typically use supervised deep learning to map images to manually annotated lesion masks. While the performance on curated datasets can be strong, generalization to unseen lesion types and different scanners remains limited^{25;78}, and assembling large expert-annotated datasets is costly²³. This tension underscores the need for methods that reduce dependence on manually created lesion maps. Such approaches must not only generalize across diverse lesion types, sizes, and contrasts but also adapt to the wide range of individual brain anatomies and disease presentations observed in clinical populations while remaining robust to demographic variability and nuisance variables.

Unsupervised Anomaly Detection (UAD) has emerged as a widely adopted alternative to supervised approaches in MRI analysis and lesion detection. The central principle of UAD is to model a representation of normal brain structure using clean data and to flag deviations from this learned concept of normality as potential anomalies in unseen cases⁶³. By shifting the focus from predefined lesion labels to deviations from normative patterns, UAD provides a framework that can potentially capture unexpected, subtle, or previously uncharacterized pathologies, making it particularly valuable for heterogeneous and poorly understood brain disorders. In brain MRI, recent deep UAD methods typically fall into two main categories: (i) *reconstruction-based* approaches, which learn to compress and reconstruct healthy images and then use voxel-level residuals between the input and reconstruction to localize anomalies^{11;22;48}, and (ii) *feature-based* approaches, which extract intermediate representations from neural networks and apply a secondary detection model for pixel-level anomaly localization^{20;62;79}. Furthermore, training strategies based on synthetic-anomalies have been introduced to further reduce reliance on real lesion maps by altering images of healthy individuals. Often, these approaches incorporate *reconstruction-based* methods to mitigate either detection or generalization problems^{59;80}.

Analogous unsupervised anomaly detection techniques have already been established in other domains, such as industrial surface defect detection^{12;20;62;79}, autonomous driving³⁰, and time series analysis^{4;13}, where standardized datasets and evaluation protocols have catalyzed progress. In contrast, progress in medical imaging has long been hampered by the lack of comparable large-scale resources. Only with the advent of initiatives such as the Human Connectome Project^{14;70;74} and UK Biobank imaging⁵⁷ has this begun to change. Perhaps unsurprisingly, as medical data is inherently sensitive, it is more difficult to acquire, share, and pool at scale⁶⁰. To compensate, some studies have resorted to practices utilizing a “healthy” reference from clean slices of anomalous volumes. This strategy potentially introduces systematic biases and undermines the generalizability and robustness of findings, as non-anomalous slices

in brains with lesions can hardly be considered to belong to a healthy cohort of brains¹⁰. Beyond the challenge of reference definition²⁹, evaluations in medical imaging often differ in fundamental aspects, including performance metrics, lesion taxonomies, brain regions analyzed, preprocessing strategies, and the degree of control for scanner and demographic factors that render the comparison of methods impossible. Although valuable steps toward unification have been made^{8;45}, these efforts remain fragmented and leave critical aspects inconsistently addressed, such as evaluation with healthy data, the diversity of the lesions tested, the impact of demographics on predictions, and missing thresholding strategies to derive the predictions. As a result, a stringent and clinically grounded benchmark is needed to establish the current state of the field, create a fair basis for comparison, and guide the development of more robust and generalizable methods in the future.

To address this need, we introduce a comprehensive, multi-center benchmark of deep UAD for brain MRI (see Fig. 1). Our training cohort is five times larger than that in previous studies and includes scans across a large age range and both sexes. Evaluation spans a holdout set of healthy individuals, diverse lesion types, and complex diseases while explicitly assessing the effects of different scanners and demographics (in particular, age and sex). Alongside leading and established UAD approaches developed for medical image analysis, we include industrial surface defect detection methods to test cross-disciplinary insights and enable direct comparisons of state-of-the-art techniques across fields. Our evaluation of stroke, tumors, multiple sclerosis (MS), and white matter hyperintensities (WMH) demonstrates that *reconstruction-based* approaches, many of which are based on diffusion models, achieved the strongest overall performance, while *feature-based* methods showed greater robustness to distributional shifts. At the same time, systematic biases linked to lesion size, image contrast, age, sex, and scanners remain major obstacles to clinical translation.

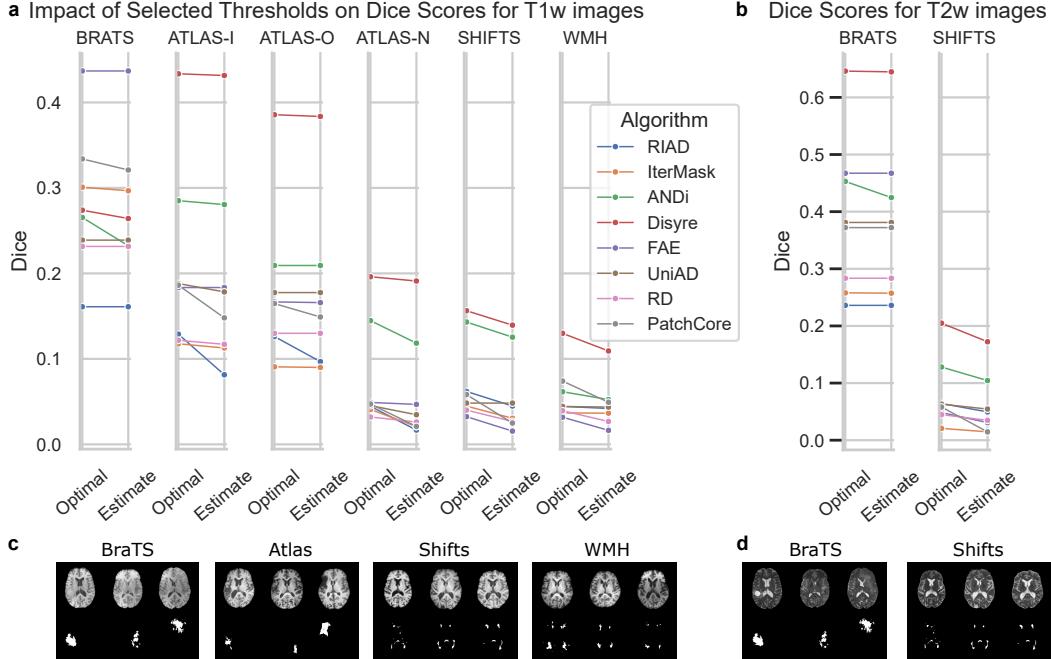


Figure 2: Multi-site and multi-task evaluation of state-of-the-art unsupervised anomaly detection

a) The performance of the algorithms on T1w images were reported using two different thresholds, the *optimal* threshold, defined as the maximum possible Dice score optimized in the test set, thus potentially susceptible to bias but standard in the field. Second, the *estimated* threshold optimized on the validation data set, then fixed and performance for that threshold reported on the untouched test set, thus unbiased but not the standard in the field. For the large lesions we did not observe a substantial difference between the thresholding procedures. **b)** The performance of the algorithms with the *optimal* and *estimated* thresholds on structural T2 weighted images. **c-d)** Example images and labels for each dataset and modality. Taken together, we show heterogeneous performance across lesions and modalities, no single method was superior across tasks. Across all evaluations the best performing methods were Disyre followed by ANDi.

2 Results

We trained six different models belonging to the broader category of *reconstruction* and *feature-based* methods. All models described in Sec. 4.3 were trained separately using T1-weighted (T1w) and T2-weighted (T2w) MRIs. We used 2976 T1w scans and 2972 T2w from healthy individuals, which amounts to approximately 461,000 imaging slices aggregated from multiple datasets for training. Details regarding the data are described in Tab. 1 and Secs. 4.1 and 4.2. After training, the models represent a normative con-

cept that can be used to calculate deviations at the voxel level and are referred to as anomaly maps.

The calculated anomaly maps are continuous, and a threshold τ is needed for the binary classification of a voxel as either healthy or anomalous. In the literature, this threshold is usually not estimated^{11;22;48;55}, and the test set, which typically contains only a single lesion type, is used to determine the *optimal* τ , potentially leading to inflated benchmarks⁸³. In a clinical scenario, this corresponds to instances in which the condition is already known. By contrast, we estimated the threshold by using a separate validation

dataset that contains all lesion types observed during the test phase, including tumors, chronic stroke, multiple sclerosis (MS), and white matter hyperintensities (WMH). Specifically, the threshold is tuned for each method by finding the best mean Dice score, which results in an unbiased threshold based on optimal performance in the validation set concerning lesion detection. Note that tuning the hyperparameters or the threshold using anomaly data is in contrast to unsupervised learning, but is the standard approach for deep learning methods in anomaly detection⁶⁹.

We evaluated all implemented algorithms with the *optimal* and *estimated* thresholds on a highly heterogeneous collection of lesions spanning multiple pathologies, sizes, contrasts, and distributions, and we refer the reader to Sec. 2.1 for the results. In addition to the evaluation of volumes containing lesions, the performance of the algorithms at the *estimated* threshold was evaluated on hold-out healthy data in Sec. 2.2. In Sec. 2.3, the effects of domain shifts, lesion size, and their interaction on detection performance were examined. In Sec. 2.4, the effects of demographic variations were systematically evaluated. Lastly, the influence of data abundance for reference definition has been examined in Sec. 2.5. The concept of the benchmark and the categories of the methods used are shown in Fig. 1. For more details regarding the individual methods and the data used, we refer the reader to Sec. 4.

2.1 Algorithmic Performance Across Brain Lesions

Among all evaluated methods, Disyre⁵⁹ achieved the highest overall performance, followed by ANDi²² and FAE⁵⁵ (see Fig. 2). In general, algorithms performed better on larger lesions, on hyperintense lesions (e.g., in T2w images), and on in-distribution (ID) samples. By contrast, *feature-based* methods consistently struggled to detect small lesions such as MS and WMH, as well as the predominantly small lesions in the ATLAS-N dataset. Interestingly, the same methods appeared more robust to out-of-distribution (OOD) samples in ATLAS-O, suggesting a trade-off between sensitivity to small, subtle lesions and ro-

bustness to distributional shifts. Threshold selection emerged as a critical factor for performance, particularly for smaller lesions (MS, WMH, ATLAS-N), indicating that deviations are shaped not only by lesion type but also by lesion size. For larger lesions, threshold sensitivity was more limited and was primarily observed in a subset of algorithms, such as PatchCore, ANDi, and RIAD. This finding underscores the importance of realistic evaluations and suggests that adaptive thresholding strategies based on the specific individual or their demographics could improve future methods.

2.2 Algorithmic Performance Across Healthy Variations

In classical anomaly detection, the maximum acceptable false positive rate is predefined, and the threshold is chosen accordingly^{27;47}. Here, we optimized the threshold using the Dice score on the validation dataset, as it balances precision and recall and is the standard metric for medical image segmentation⁵⁸. Since different lesion types (tumors, stroke, MS, WMH) each have distinct ideal balances with the negative class (healthy voxels), we used a validation dataset containing all lesion types to define an unbiased operating point with meaningful detection accuracy and low false positive rates. We evaluated the resulting false positive rate in the healthy cohort to understand how the realistic operating point influences false alarms for healthy individuals under domain shifts that could not have been accounted for with the ideal balance in the validation dataset. Three datasets containing only healthy individuals, imaged with different protocols, were used to evaluate the false positive rate of the algorithms. Note that these datasets represent the general population, where lesions are extremely rare, and samples are generally lesion-free. Applying our algorithms to such data, therefore, provides a good estimate. As shown in Fig. 3, all algorithms performed differently across the three distinct datasets, indicating a strong influence from either imaging protocols or demographics. Overall, the methods showed the highest false positive rates (FPRs) on the 1.5T FCON scans and the best performance on TCP. We attribute the reduced

performance on FCON-1.5T to lower image quality and the limited number of 1.5T scans in the training data, while FCON-3T likely includes OOD samples from unseen scanners or protocols. TCP scans belong to the ID distribution, which could explain the higher performance. In general, FPRs in healthy samples correlated with lesion detection performance. Disyre performed best, followed by ANDi and PatchCore. Interestingly, some *feature-based* approaches (e.g., PatchCore, UniAD) achieved lower FPRs but exhibited weaker detection, and their performance varied significantly across datasets, indicating bias from other factors. For T2w scans, FPR values were computed on TCP (Suppl. Fig. S1). Importantly, this analysis highlights threshold selection as a key bottleneck for clinical translation: despite using a diverse validation dataset, the thresholds produced large FPR differences across healthy datasets.

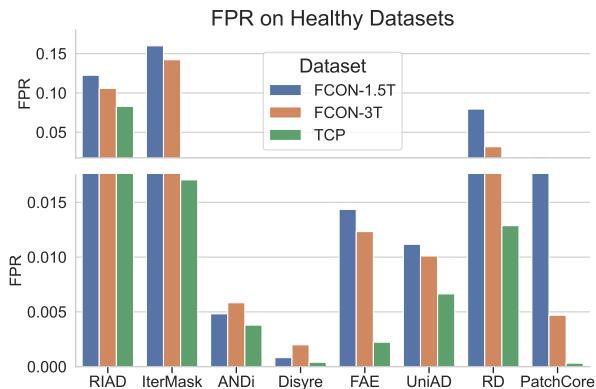


Figure 3: False positive rate on healthy brains across all tested algorithms associated with estimated threshold. For some methods we identified high false positive rates and heterogeneous performances across the healthy cohort for T1w images (for T2w see supplement). This shows that the methods were biased for either imaging protocols or demographics and selecting the right decision threshold is critical.

2.3 Impact of Domain Shifts and Lesion Load Variability

Robustness to domain shifts, particularly those induced by differences in MRI scanner hardware, is essential for clinical translation. Using the ATLAS dataset, we partitioned the data into ID and OOD subsets, stratifying cases by lesion load, i.e., number of anomaly voxels in the volume, to investigate how lesion burden interacts with domain variability (see Fig. 2 and Sec. 2.1). We constructed eight datasets corresponding to four matched ID–OOD pairs, stratified by lesion load percentiles derived from the full ATLAS cohort. We group volumes above the 75th percentile, those above the 50th percentile to the 75th, those above the 25th percentile to the 50th percentile, and those below the 25th percentile. This design ensured that each pair contained scans with comparable lesion burdens, thereby isolating the effects of scanner heterogeneity from lesion size. Statistical evaluation using Mann–Whitney U tests with Benjamini–Hochberg correction revealed a significant decline in segmentation performance for most *reconstruction-based* methods when applied to OOD data with high lesion burden, while *feature-based* methods remained largely unaffected, as shown in Fig. 4. The only exception was FAE, which showed significant performance losses in the upper and middle lesion load percentiles. Across all methods, Dice scores decreased with decreasing lesion load, where lesions are smaller and more subtle, thereby challenging algorithmic sensitivity, and the performance gap between ID and OOD samples narrowed for smaller lesions, indicating that lesion size exerts a stronger influence on model performance than scanner-related factors. Nevertheless, even top-performing algorithms displayed marked drops in the top and upper lesion load strata under domain-shift conditions, exposing critical vulnerabilities to scanner differences and lesion variability. The p-values and corrected p-values can be found in Suppl. Table 6. To further test the interaction between lesion load and domain variability, we conducted a Scheirer-Ray-Hare test with the lesion load categories and the ID–OOD dummy variable as the independent variables to predict the Dice scores. All algorithms were significantly

Dice Scores for In Distribution vs Out of Distribution split by Lesion Load

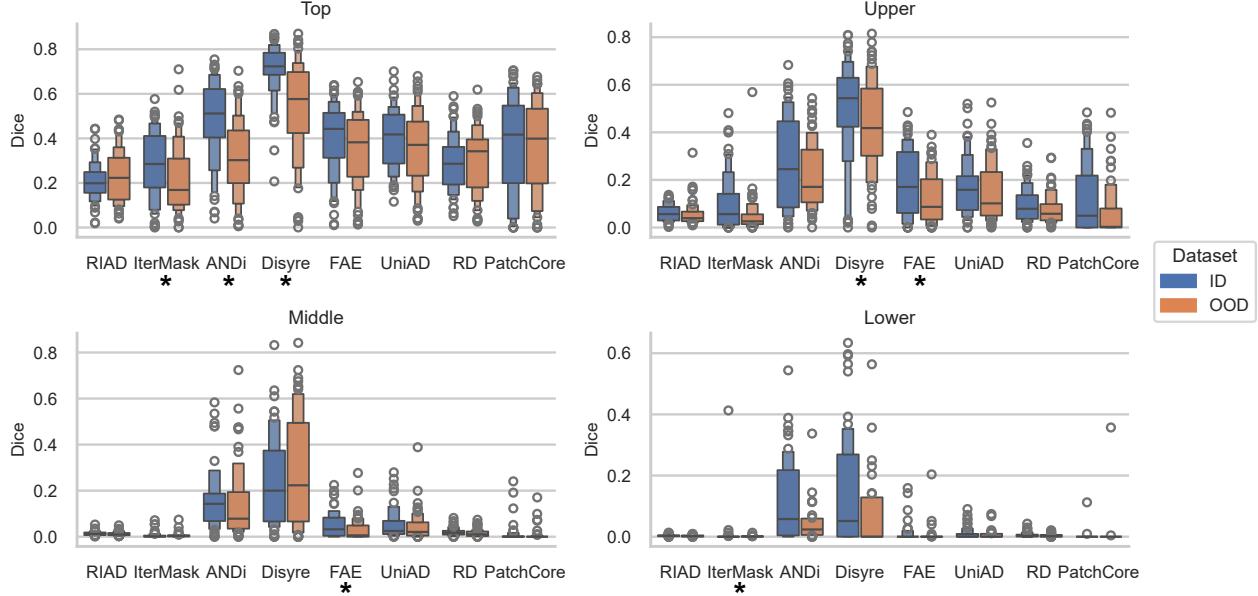


Figure 4: **Out-of-distribution effects due to scanner differences and influence of lesion load.** The ATLAS dataset had been split by the percentiles of the lesion load. Top corresponds to a lesion load above the 75th percentile, upper - above 50th percentile to 75th, middle - above 25th percentile to 50th percentile, lower - below 25th percentile. Columns marked with * correspond to the distributions of Dices scores that are different according to the Mann-Whitney U test after multiple testing correction with the Benjamini–Hochberg adjusted significance level $\alpha = 0.05$. Taken together, *reconstruction-based* approaches are more sensitive to imaging protocols, while *feature-based* methods are generally robust to this source of variation in the data.

impacted by lesion load; ANDi and Disyre were impacted by the ID-OOD split, and Disyre exhibited a significant interaction term. The degrees of freedom and p-values for the Scheirer-Ray-Hare tests can be found in Suppl. Table 7. These results underscore the urgent need for models that generalize reliably across acquisition settings and patient populations, particularly in scenarios with small, clinically relevant lesions.

2.4 Impact of Age and Sex on false positive rates and Lesion Identification

The performance of unsupervised anomaly detection methods varied substantially across individuals due to multiple interacting factors. Lesion size was a key determinant of detection accuracy, with larger lesions producing stronger deviations and being more readily detected, while false positive rates showed considerable variation across methods and datasets (see Fig. 3). To further analyze this behavior, we studied the impact of inter-individual anatomical variability linked to demographic attributes such as age and sex in the HCP Aging and BraTS datasets. As shown

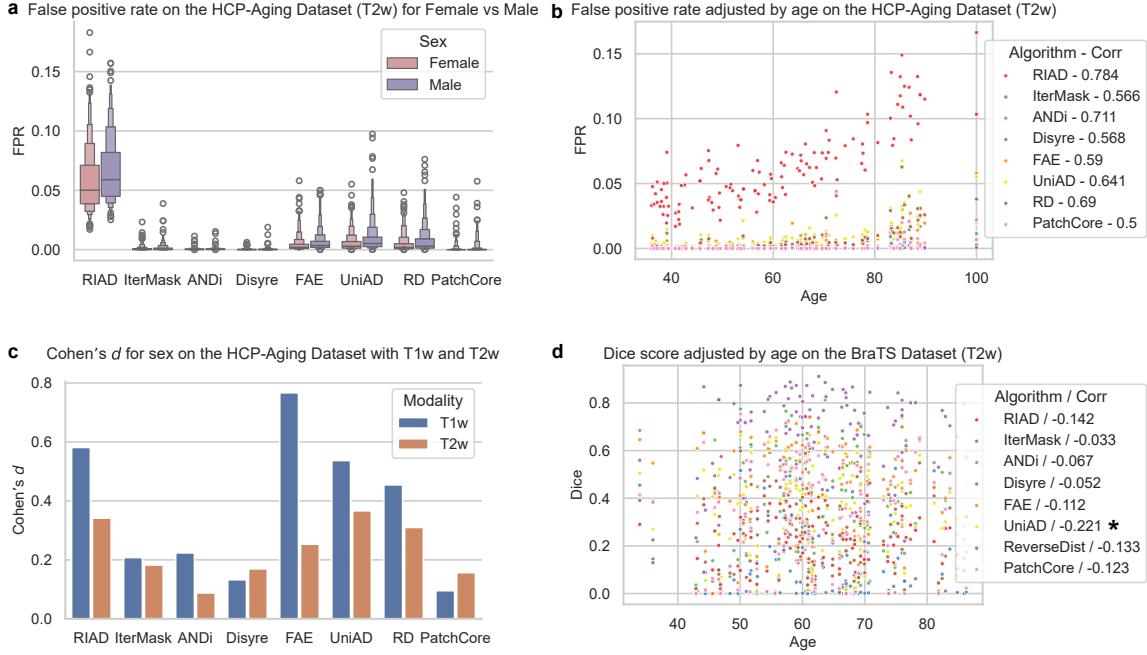


Figure 5: **Impact of demographics on false positive rates and lesion identification.** **a)** The false positive rate on the HCP Aging dataset for the female and male group. All groups show significant differences based on the Mann-Whitney U test. Significance level $\alpha = 0.05$ was used for all tests. **b)** The false positive rate on the HCP Aging dataset adjusted by age. The Spearman rank correlation test displayed that all algorithms show significant positive correlation indicating that older individuals are more likely to have an anomaly assigned to a voxel that is normal. **c)** Cohen's d on the HCP Aging dataset for the groups male and female with both modalities. All algorithms show a positive Cohen's d , an effect due to higher false positive rates for males. **d)** The Dice score on the BraTS dataset adjusted by age. Only UniAD (marked with *) shows a significant negative correlation (Spearman rank correlation). Bias for age and sex on healthy volumes generally decreased with increasing performance of the algorithm.

in Fig. 5, the Spearman rank correlation test with $\alpha = 0.05$ indicated that models tended to overestimate the occurrence of anomalies in older individuals in the HCP Aging dataset, likely due to age-related changes in brain structure. Interestingly, there was no effect of age on the Dice score. To evaluate differences between the sexes, we calculated a Mann-Whitney U test on the FPR as a proxy for the fairness of our algorithms. All algorithms showed significant differences in performance for the different sex groups, with male brains being predicted as more anomalous than female brains. Most algorithms ex-

hibited higher effect sizes on the T1w images, as measured by Cohen's d , indicating stronger sensitivity to sex-related differences in this modality. Importantly, the top-performing algorithms were among those with the lowest bias, suggesting that high overall accuracy can imply, in our scenario, better fairness across biological groups. Detailed results, including the means and standard deviations of the FPR values for both groups, the corresponding Cohen's d values, and the full plots for the T1w analyzes, are provided in Suppl. Fig. S2. These findings highlight the need to develop and rigorously validate anomaly detection

models that are robust against biological variability. Such resilience is critical not only for the reliable detection of lesions but also for ensuring stable performance in healthy cohorts, thereby supporting both clinical applicability and generalizability across diverse populations.

2.5 Effect of Data Scaling During Training

To assess the influence of data abundance on performance, we tested the impact of diminished training data and experimented with an additional training dataset that comprised 3% of the described training dataset used in the main analysis. For the performance evaluation of the newly trained models in the detection task, only the validation dataset was used due to computational constraints. As shown in Fig. 6, all algorithms except UniAD performed almost identically in terms of the mean Dice score on the validation dataset. Similarly, performance on false positive rates measured with the TCP dataset was stable across training datasets, with two exceptions: UniAD and IterMask. These results suggest that straightforward scaling strategies alone, by enlarging training data, are insufficient to overcome fundamental limitations for most unsupervised anomaly detection tasks. The analysis was performed on the T1w images, and RIAD was excluded due to difficulties encountered during the optimization procedure for all conducted experiments.

3 Discussion

This study presents one of the most comprehensive evaluations of UAD for brain MRI to date, spanning multiple scanners, demographic groups, and four major brain diseases, ranging from stroke and tumors to multiple sclerosis (MS) and white matter hyperintensities (WMH). While recent advances demonstrate clear progress, current approaches remain far from achieving the robustness, sensitivity, and fairness required for clinical deployment⁴⁰. *Reconstruction-based* methods, particularly diffusion-inspired approaches such as Disyre and ANDi^{22;59}, achieved the

highest segmentation accuracy, especially for large and hyperintense lesions. *Feature-based* methods, by contrast, were more resilient to scanner variability but consistently struggled with subtle or small anomalies. Several newly developed *reconstruction-based* methods have achieved state-of-the-art performance, but no algorithm achieves clinically relevant performance defined by sufficient detection performance, robustness to domain shifts caused by imaging protocols and potentially new populations, and without unintended discriminatory biases⁴⁰. On the contrary, our findings highlight systematic sources of bias, namely that lesion size and image contrast were dominant drivers of accuracy, and false positive rates in healthy cohorts were systematically influenced by demographics, particularly age and sex. Taken together, these results emphasize that future advances must go beyond incremental performance gains. Methodological priorities include increasing sensitivity to small and low-contrast lesions, improving robustness to distributional shifts across scanners and protocols, reducing demographic biases, and refining evaluation procedures to more closely mirror real-world conditions.

A central theme that emerges from this study is the critical role of thresholding and evaluation practices. Because UAD produces continuous anomaly maps, binarization requires a decision threshold. While threshold-independent evaluation exists, it does not provide an indication of the performance under a specific binary decision that is needed for any future diagnosis or longitudinal monitoring; therefore, it is less valuable in a medical context. In much of the literature, thresholds are optimized directly on the test set, potentially resulting in inflated performance estimates that would not hold in clinical settings. In contrast, we determined thresholds using a heterogeneous validation set encompassing multiple lesion types and then evaluated them on unseen test data. This more principled procedure yielded clinically realistic performance estimates that showed differences from standard thresholding procedures for some of the compared models (see Fig. 2). Such effects underscore that thresholds are not simple post-hoc technicalities but integral components of the detection pipeline. Effective evaluations and eventual deploy-

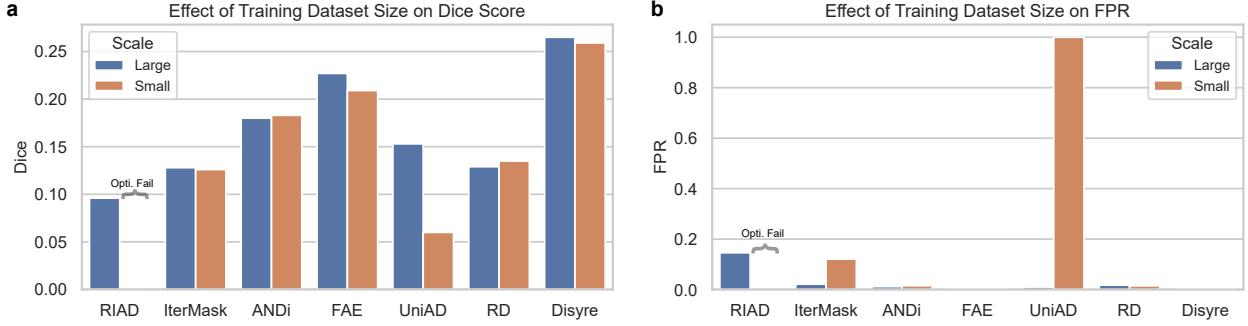


Figure 6: Impact of dataset scaling on algorithmic performance **a)** The Dice score on the T1w validation dataset. **b)** The false positive rate for the T1w images in the TCP dataset. The training scale is defined as the number of individuals in the dataset. The small training scale uses 3% (92 volumes) of the training dataset. This analysis showed that the majority of algorithms are irresponsive to more data, highlighting the need for methods that are capable to grain performance when integrating large scale data for anomaly detection.

ment, therefore, require adaptive thresholding strategies potentially conditioned on scanner, protocol, or demographic information, as well as standardized validation procedures to ensure fair and reproducible comparisons across methods.

Closely related to this issue are the implicit assumptions underpinning *reconstruction-based* approaches, which we deem to be a major cause of their shortcomings with respect to clinical usability. The first assumption is that a model trained exclusively on healthy data will always reconstruct a healthy sample. While diffusion-based methods partly relax this assumption through the forward process, enabling a tradeoff between reconstruction fidelity and an image from the data distribution, interference in this process is practically non-trivial^{9;22}. Furthermore, control of the reverse process to create conditional samples remains mathematically opaque for diffusion and flow models⁵⁰. More classical methods, such as RIAD or IterMask, employ complex masking strategies to encourage useful reconstructions; however, these do not guarantee that anomalies are removed. Even subtle intensity perturbations can suffice when residual error is used. The second assumption is that the deviation measure between input and reconstruction is inherently suitable for anomaly localization. The commonly used residual error has

well-documented drawbacks⁵⁴, as it treats each pixel independently and ignores spatial context. Alternatives such as the Structural Similarity Index Measure (SSIM), as used in the FAE⁵⁵, partly address this by incorporating locality. However, no deviation metric or measure has yet been explicitly designed for brain MRI. Therefore, we suggest that the choice of deviation is as critical as the reconstruction itself, and developing neuroanatomically informed measures could represent a key avenue for advancing *reconstruction-based* UAD.

In contrast to previous reports⁴⁵, *feature-based* approaches did not achieve the highest performance in our benchmark. Consistent with earlier analyzes, however, we observed that these methods were particularly weak at detecting small lesions. We hypothesize that this relative decline reflects the rapid advances in *reconstruction-based* and generative modeling approaches, which have shifted the comparative performance landscape^{20;22;48;55;59;62;79;81}. The central assumption underlying *feature-based* frameworks is that pretrained features provide sufficient representational power to distinguish anomalies from normal tissue in the embedding space. All *feature-based* methods included here relied on ImageNet-pretrained networks, even when originally developed for brain MRI (e.g., FAE). While such features can yield com-

petitive performance, the use of ImageNet pretraining is suboptimal, and the extension to MRI-specific pretraining appears to be a natural progression. Indeed, prior work suggests that MRI-based pretraining can yield small but consistent improvements⁴⁵. We did not pursue this option in the present benchmark, as no recent UAD method has demonstrated state-of-the-art performance using MRI-pretrained features. The scarcity of such approaches in the literature likely reflects both practical and conceptual barriers. One important barrier is the lack of MRI datasets that are comparably large and harmonized to the scale of ImageNet¹. Another difficulty is the design of pretext tasks that transfer effectively to the downstream challenge of anomaly detection. Overcoming the former barrier is likely infeasible, but the latter could reinvigorate *feature-based* pipelines and potentially restore their competitiveness, especially if paired with strategies that address scanner heterogeneity and small-lesion sensitivity.

Sensitivity to domain shifts caused by scanner effects emerged as an important take-home message as a result of our benchmark. By stratifying the ATLAS dataset into in-distribution and out-of-distribution subsets matched for lesion load, we were able to disentangle the effects of scanner variability from lesion size. *Reconstruction-based* methods showed pronounced degradation on OOD data, particularly when lesion burdens were high, suggesting that larger lesions may amplify reliance on scanner-specific statistical cues. *Feature-based* methods were generally more stable across scanners. However, exceptions such as the FAE model highlight how deviation measures strongly influence generalization. We suspect that the FAE’s susceptibility to OOD shifts arises from its use of an image-derived deviation measure, which treats high-resolution feature maps as if they were natural images and therefore inherits scanner sensitivities. Importantly, our results reveal that lesion size and domain shifts can interact, with large lesions amplifying vulnerability to acquisition-related biases, whereas small lesions are inherently more difficult to detect. These findings emphasize that robust generalization to heterogeneous imaging protocols remains an unsolved challenge for any possible future deployment. Address-

ing this issue will require explicit domain adaptation strategies, including scanner-aware harmonization⁷⁶, training on more diverse acquisition protocols, and potentially test-time adaptation to local distributions to ensure consistent performance across imaging environments. Demographic factors also had a systematic effect. False positives increased with age in healthy cohorts, reflecting normal age-related changes misclassified as pathology. All algorithms showed significant sex-related differences, with male brains more often flagged as containing anomalous regions. Although the top-performing methods exhibited somewhat reduced demographic bias, these results underscore the need for fairness-aware modeling¹⁰. Incorporating demographic factors into normative reference frameworks⁶⁴ and anomaly detection workflows, explicitly quantifying subgroup biases⁶⁵, and potentially calibrating thresholds in a stratified manner will be essential to ensure equitable performance. More broadly, these observations point to the necessity of developing models that are resilient to natural anatomical diversity and capable of distinguishing genuine pathology from normative variability.

We observed a limited impact of scaling training data. Whereas supervised machine learning typically benefits substantially from larger datasets^{21;31}, simply adding more healthy scans yields only minor improvements in UAD performance for some algorithms. This reflects the algorithmic constraints of the UAD approaches, for which training solely on healthy cohorts cannot effectively teach models which deviations are clinically meaningful. Additionally, subtle lesions are often indistinguishable from normal anatomical variation. In line with prior work⁴⁵, our experiments confirmed that even when the training data were drastically reduced, most methods showed comparable lesion detection accuracy and false positive rates for in-distribution evaluations. These results suggest that naive scaling is insufficient to overcome the intrinsic challenges of neuroimaging anomaly detection. Future progress will require qualitatively new strategies, including anatomically informed deviation measures, large-scale medical pre-training with task-aligned self-supervised objectives, and robust domain adaptation frameworks that explicitly account for clinical variability.

Although our study drew on a broad range of datasets spanning multiple lesion types and imaging protocols, the collection cannot fully reflect the diversity of real-world clinical neuroimaging. Rare pathologies⁴⁶, pediatric cohorts⁷⁰, and cases acquired under extreme or atypical imaging conditions remain underrepresented, which may limit the generalization of our conclusions. Similarly, although we explicitly examined the demographic effects of age and sex, other important axes of biological and social variability were not captured. Factors such as ethnicity, comorbidity, medication status, or socioeconomic background can shape brain anatomy and imaging characteristics⁵¹, and their influence on anomaly detection remains unexplored. Another limitation lies in our choice of evaluation targets. Thresholds were optimized with respect to segmentation accuracy, which provides a clear and widely accepted benchmark but may not align with clinical decision-making needs. Tasks such as patient-level triage, prognosis, or monitoring treatment response may require different operating points or evaluation metrics, emphasizing sensitivity, specificity, or predictive value over voxel-level overlap. Future studies should, therefore, consider multiple evaluation frameworks that reflect the range of clinical contexts in which anomaly detection could be applied. Methodological scope is another constraint. Although we implemented a broad set of state-of-the-art algorithms, not all contemporary or emerging approaches could be included. Having said that, we implemented relevant and current state-of-the-art algorithms while aiming for algorithmic diversity in the methods used. Moreover, although our pipeline was carefully designed to minimize data leakage and benchmark bias⁷⁵, it still cannot reproduce all conditions relevant to clinical deployment. In practice, anomalies are rare, heterogeneous, and often subtle, and scans are embedded in complex diagnostic workflows. These realities pose additional challenges, such as handling incidental findings, balancing false positives against clinical workload, and integrating uncertainty estimates that remain outside the scope of this study. Taken together, these limitations highlight that while our benchmark offers an important step toward the systematic evaluation of UAD in neuroimaging, further

work is needed to expand dataset diversity, incorporate broader demographic and clinical factors, refine evaluation criteria, and explore algorithmic strategies in conditions that more closely approximate real-world deployment.

Looking forward, several priorities emerge for future research. First, the design of principled deviation metrics should become a central focus. Residual errors and generic SSIM-based measures have well-documented limitations^{54;77}, and future work should aim for neuroanatomically grounded deviations. Having said that, this task is admittedly extremely difficult, and determining whether this ambition is feasible is still subject to current research^{7;19;42;43}. Benchmarks like ours, along with further developments using the principles outlined here, will allow us to evaluate such methods in more depth than is currently common. Second, large-scale medical pretraining on curated neuroimaging datasets, combined with meaningful self-supervised tasks, could provide domain-native representations that are more sensitive to subtle anomalies than ImageNet features. Third, harmonization and domain adaptation need to be built directly into the modeling pipeline to ensure robust performance across scanners and acquisition protocols. Fourth, fairness-aware modeling should be prioritized¹⁰, with systematic evaluation of demographic biases and strategies to mitigate them. Finally, the evaluation itself must incorporate thresholds, as outlined here, and the reporting of uncertainty, robustness, and fairness alongside accuracy is essential to establish clinical trust.

In conclusion, this benchmark underscores both the promise and the current limitations of UAD in brain MRI. While modern approaches can detect a broad range of lesions, their performance is uneven across lesion types, scanners, and populations. They remain susceptible to bias and thresholding procedures. Addressing these challenges will require innovations that go beyond scaling data or marginal architectural tweaks. By developing principled deviation metrics, MRI-native pretraining, robust domain adaptation, fairness-aware pipelines, and clinically meaningful evaluation frameworks, the field can move closer to reliable, equitable, and actionable anomaly detection in neuroimaging.

4 Methods

Here, we provide information on the data (see Sec. 4.1 and Tab. 1), preprocessing (Sec. 4.2), and the evaluated methods (see Sec. 4.3 and Tab. 2). Additional information about training and tuning for all methods can be found in the supplementary methods.

4.1 Datasets

For training, we used large-scale healthy cohorts, including the Cambridge Center for Aging and Neuroscience dataset (CAMCAN)⁷², the Human Connectome Project (HCP) Young Adult (S1200) dataset⁷⁴, the HCP Development dataset⁷⁰, and the IXI dataset⁵², resulting in 2,976 T1w and 2,972 T2w scans acquired across six scanners. For validation, we randomly selected approximately 10% of the individuals from all other lesion datasets, described below, yielding about 92 T1w and T2w scans. This dataset was used to tune hyperparameters and estimate unbiased thresholds. For testing, we included both healthy and clinical cohorts. The healthy cohort comprised the Transdiagnostic Connectome Project (TCP)¹⁶, a subset of the 1000 Functional Connectomes Project (FCON)⁶⁸, and the HCP Aging dataset¹⁴. The clinical cohort was made up of the Multimodal Brain Tumor Segmentation Challenge (BraTS 2020)^{5;6;56}, the Anatomical Tracings of Lesions After Stroke (ATLAS v2.0)⁴⁹, the MSSEG and Ljubljana MS lesion datasets from the 2021 Shifts Challenge⁵³, and the White Matter Hyperintensities (WMH) Challenge dataset from MICCAI 2017⁴⁴. In total, the test set consisted of 2,221 T1w and 1,262 T2w scans, with the distribution of lesions and healthy samples summarized in Tab. 1 and visualized in Fig. 1. To analyze scanner-related domain shifts, the FCON dataset was divided into 139 scans acquired on 1.5T scanners (München, Oulu, Orangeburg) and 105 scans acquired on 3T scanners (Atlanta, Palo Alto, Ann Arbor). The ATLAS dataset was further partitioned into AtlasI (262 scans from scanners also used during training), AtlasO (227 scans from unseen scanners), and AtlasN (166 scans from scanners with unidentifiable information, which by chance contained mostly smaller lesions). Addi-

tionally, the full ATLAS cohort was stratified into four groups (Top, Upper, Middle, and Lower) based on lesion load, using the 25th, 50th, and 75th percentiles, and combined with the scanner-based partitions, resulting in 57 individuals for ID Lower, 34 for OOD Lower, 51 for ID Middle, 58 for OOD Middle, 77 for ID Upper, 58 for OOD Upper, 79 for ID Top, and 77 for OOD Top. These splits allowed us to systematically evaluate algorithm robustness under distributional shifts induced by both scanner variability and lesion burden.

4.2 Preprocessing

For preprocessing the CAMCAN, IXI, FCON, TCP, ATLAS, and WMH datasets, we applied a pipeline similar to the UK Biobank protocol². Using FSL³⁷, all images were reoriented to match MNI152 space²⁸ (rotation only, no registration). For multimodal datasets, rigid transformations to the individual’s T1w image were computed with FLIRT³⁶ and applied to align the modalities. This ensured that all subsequent transformations could be calculated on the T1w and transferred to other contrasts. Because Siemens gradient nonlinearity correction requires proprietary files and not all scans were Siemens acquisitions, this step was omitted. Instead, preprocessing proceeded with a field-of-view reduction on the T1w images, as in the UKB pipeline. Non-linear registration to MNI152 space was performed with FNIRT³ to generate a standard-space brain mask, which was then inverted and applied for skull stripping across modalities. After skull stripping, all scans were rigidly registered to the SRI24 atlas⁶¹ and resampled to 1 mm isotropic resolution. In cases where large anomalies (e.g., ATLAS lesions) rendered FNIRT unreliable, ROBEX³⁴ was used for skull stripping. For HCP datasets, we used the minimally preprocessed scans²⁶, additionally applying rigid registration and interpolation to SRI24 at 1 mm isotropic resolution for consistency with other datasets. Shifts challenge data⁵³ was already denoised, skull stripped, bias-field corrected, and interpolated; we therefore only applied rigid registration to SRI24. One Ljubljana case with a failed skull stripping was excluded. BraTS scans had un-

Table 1: MRI Details of Multi-Side Data

Dataset	# Scans	# Scanners	Modalities	Condition
CAMCAN	652	1	T1w, T2w	Healthy
HCP 1200	1088	1	T1w, T2w	Healthy
HCP Dev	651	1	T1w, T2w	Healthy
IXI	580	3	T1w, T2w	Healthy
TCP	92	1	T1w, T2w	Healthy
FCON	244	Unknown	T1w	Healthy
HCP Aging	725	1	T1w, T2w	Healthy
BraTS	369	Unknown	T1w, T2w	Tumor
Atlas	655	14	T1w	Stroke
Shifts	76	Unknown	T1w, T2w	MS
WMH	60	3	T1w	WMH

dergone skull stripping, rigid registration to SRI24, and interpolation, but with differing orientations; these were reoriented with FSL. It is important to note that the dataset collection included both bias-field-corrected and uncorrected scans, as well as varying skull-stripping procedures across training and test sets. For deep learning preprocessing, all volumes were min–max normalized to [0,1], converted into axial slices, and zero-only slices were removed. The remaining slices were center-cropped to 224×224 . Algorithm-specific preprocessing exceptions are described in the Supplementary Methods.

4.3 Models, Architectures and Training

The *reconstruction-based* approaches used in this study include Reconstruction-by-Inpainting Anomaly Detection (RIAD)⁸¹, Iterative Spatial Mask-Refining (IterMask)⁴⁸, Aggregated Normative Diffusion (ANDi)²² and Diffusion-Inspired Synthetic Restoration (Disyre)⁵⁹. We parameterized all *reconstruction-based* approaches with the same U-Net architecture that was inspired by the DDPM++ model from⁷¹.

RIAD⁸¹ is a self-supervised approach trained to predict masked regions of varying sizes within an image. The image is first partitioned into regions of size

$k \times k$, each of which is randomly divided into n disjoint subsets that serve as the masking patterns. For each subset, the network predicts a reconstruction, and the n reconstructions are combined to yield an image in which every pixel has been estimated by the model. This process is repeated multiple times while varying the parameter k , thereby enforcing predictions at multiple spatial scales. The reconstruction losses across scales are averaged using the multi-scale gradient magnitude similarity measure, which encourages fidelity to structural features. The key idea is to obscure potentially anomalous regions while enabling the network to generalize across diverse lesion types through complex multi-scale masking strategies. In our experiments, we increased k to ensure coverage of large brain lesions while proportionally increasing n , thereby maintaining sufficient contextual information for accurate reconstruction.

IterMask⁴⁸ is a self-supervised approach that utilizes two models trained with multiple masking strategies to predict the original image. One model is trained using the masking of low-frequency components in an image. The other model is trained by additionally using randomly generated masks in pixel space. Consequently, both models receive the high-frequency components of the image as input, while the second one is able to integrate the complete frequency information from random image locations for

Table 2: Origin and Framework of the Methods

Method	Origin	Framework
RIAD	Industry	Reconstruction
IterMask	Neuroimaging	Reconstruction
ANDi	Neuroimaging	Reconstruction
Disyre	Neuroimaging	Reconstruction + Synthetic Anomalies
FAE	Neuroimaging	Feature-based with SSIM
UniAD	Industry	Feature-based
RD	Industry	Feature-based
PatchCore	Industry	Feature-based

reconstruction. Both models are used to create an iterative process that starts by first predicting the image from its high-frequency components using the first model, and then iteratively applying the second model on the resulting mask from the previous step, aiming to shrink the mask towards the anomalies. At each iteration, a new mask is generated by thresholding the reconstruction error from the previous step. The threshold is determined on a healthy validation set, and the iterative process is terminated once the relative change between consecutive steps falls below a predefined ratio.

ANDi²² is a diffusion model^{32;33;41} trained using Gaussian pyramidal noise, which injects perturbations at multiple spatial scales and thereby enhances sensitivity to low-frequency structures. This training strategy allows the network to effectively reason about broad, slowly varying anomalies throughout the entire denoising trajectory, rather than focusing solely on high-frequency details in the early time steps. To compute anomaly evidence, ANDi evaluates a selected subset of diffusion time steps. At each step, the predicted mean of the Gaussian transition is compared against the image-conditioned mean, and the squared reconstruction error is used as the local deviation measure. This step-wise evaluation captures discrepancies that may emerge at different stages of the diffusion process. To derive a single anomaly map, deviations across timesteps are aggregated using the geometric mean, a choice that emphasizes consistent deviations across scales while attenu-

ating spurious errors occurring at isolated steps. The resulting anomaly map thus integrates information across multiple diffusion time steps and frequency ranges, offering a principled estimate of abnormality that is both spatially localized and robust to noise.

Disyre⁵⁹ is inspired by diffusion models and employs synthetic anomalies to construct a corruption process. Anomalies are generated using a novel Disentangled Anomaly Generation (DAG) framework, which independently samples shape, texture, and intensity attributes from uniform distributions. The shape of each anomaly is created by randomly selecting cuboids, spheres, or other 3D primitives, which are further modified using affine transformations and smoothed with a Gaussian kernel to blend with the surrounding tissue. Texture is defined through Foreign Patch Interpolation (FPI), where a random patch from the training set is inserted into the target image using a convex combination with a randomly sampled interpolation factor. The patch is normalized before insertion to prevent confounding between texture and intensity. Intensity is determined by sampling a bias factor and a tissue class identified through k-means clustering, and then adjusting the intensities of the selected tissue within the anomaly mask according to the bias factor. In combination, these disentangled attributes govern the corruption process, and the network is trained in a diffusion-style framework to predict the original anomaly-free image, effectively learning to reverse the synthetic corruption.

Feature-based approaches included in this study were Structural Feature-Autoencoders (FAE)⁵⁵, Unified Model for Multi-class Anomaly Detection (UniAD)⁷⁹, Reverse Distillation (RD)²⁰ and PatchCore⁶². All *feature-based* approaches were parameterized by their original network or partially modified to fit the image resolution used in the experiments.

FAE⁵⁵ is a convolutional autoencoder that operates on embeddings extracted from multiple layers of a ResNet pretrained on ImageNet. The network is trained to reconstruct these embeddings using the Structural Similarity Index Measure (SSIM) as the loss function. At inference, anomaly detection is performed by computing the SSIM between the input and reconstructed feature maps, followed by averaging the similarities to produce the final anomaly score.

UniAD⁷⁹ is a transformer-based network that introduces specialized layer types for anomaly detection. They are introduced to prevent the “identical shortcut” that causes the network to simply copy the input. A key innovation is the neighbor masked attention layer, a variant of standard attention in which tokens from the local neighborhood are masked prior to the attention calculation. In addition, the architecture features a novel decoder that employs multiple query embeddings, which are fused with encoder representations and integrated with the output from the previous layer. In order for the network to learn the identity mapping, the query embeddings need to be sensitive to the input, thereby reducing the reconstruction ability for abnormal samples. Similar to the FAE, it leverages embeddings from an ImageNet-pretrained network, specifically EfficientNet; however, unlike the FAE, Gaussian noise is added to the input features, forcing the network to jointly learn reconstruction and denoising. The model is trained with an L2 loss. At inference, anomalies are detected by computing the Euclidean distance between the input and reconstructed pixel-level features.

RD²⁰ applies a knowledge distillation framework to anomaly detection. A pre-trained ImageNet encoder serves as the teacher network, while a trainable bottleneck embedding module maps the teacher’s representations into a more compact code. In conjunction, a decoder is trained to reconstruct the teacher’s em-

beddings from this bottleneck representation, with training guided by maximizing cosine similarity between the reconstructed outputs and the teacher’s original embeddings. The key idea is that, due to the heterogeneous architecture and the compression introduced by the bottleneck, the student network cannot perfectly replicate the teacher’s embeddings for novel or anomalous data. As a result, anomalies can be identified at test time by measuring the cosine similarity between the pixel-level features of the input and output, with lower similarity indicating abnormality.

PatchCore⁶² employs a memory bank of pre-processed embeddings for anomaly detection. A WideResNet-50 pretrained on ImageNet is used as the backbone, from which embeddings are extracted from intermediate layers. Prior to assembling the memory bank, the features are refined to enlarge the receptive field: overlapping patches are sampled from the feature maps, and adaptive average pooling is applied to aggregate information from each neighborhood into a single feature vector. The resulting representations form the memory bank of nominal features. To improve efficiency at inference, the memory bank is downsampled using greedy coresset subsampling. Anomaly detection is then performed by computing the mean L2-distance to the nearest entries in the memory bank. In our MRI experiments, we observed that performance improved when considering only the distance to the single closest entry, rather than averaging across multiple neighbors.

All models were trained for a maximum of 100,000 gradient update steps, and checkpoints were saved after every 5000 steps. Then, the best-performing checkpoint was chosen from the validation dataset for each method. We noticed that many methods reached optimal performance on the validation dataset before convergence.

All code, along with our adaptations of the evaluated methods into reproducible and testable workflows, is available on GitHub (<https://github.com/AlexanderFrotscher/UAD-IMAG>) and has been forked to the MHMlab repository (<https://github.com/MHM-lab>). Prior to submission, we contacted the lead authors of the original papers and invited them to review our implementations via the shared

GitHub repository. This process was designed to ensure transparency, fairness, and author-validated benchmarking.

4.4 Post-Hoc Statistics and Model Evaluation

After obtaining the anomaly maps, three-dimensional median filtering with a kernel size of four is applied to all algorithms. Threshold selection has been performed on the original anomaly maps and has been transferred to the median-filtered ones. For evaluating the anomaly map, many different metrics can be used; that is, all metrics that can be used for binary classification can, in principle, work for the evaluation of the UAD methods. In the medical community, two of these are now widely accepted as the standards: the Dice score and the area under the precision-recall curve (AUPRC). Note that for brain MRI, the area under the receiver operating characteristic curve is less important due to the high class imbalance between positives and negatives. Here, we mainly report the Dice score and present all AUPRC values in the Supplementary Information. Table 1 - 4. All threshold-dependent metrics have been calculated on a per-individual basis and averaged across individuals when reporting mean values, whereas the AUPRC has been calculated on the complete dataset once. To evaluate the influence of the scanner (the specific MRI device) and sex effects, the distributions of the performance measures and groups have been analyzed using the Mann-Whitney U test. It is a rank-based test that does not assume a specific parametric distribution and can indicate significant differences for all aspects of the distribution, e.g., location, scale, and shape. Multiple testing corrections using the Benjamini-Hochberg method have been applied to correct for the four datasets corresponding to the different lesion loads. Furthermore, age effects have been analyzed using the Spearman rank correlation test to assess nonlinear correlations between performance and age. For all statistical tests, a significance level $\alpha = 0.05$ has been used. To analyze the magnitude of the sex effects, Cohen's d was calculated with the male group as the first group. Therefore, all positive Cohen's d values indicate that the FPRs

were higher for the male group.

5 Acknowledgements

AF and TW thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for their support. TW acknowledges funding from the German Research Foundation (DFG) Emmy Noether: 513851350 and the BMBF/DLR Project FEDORA: 01EQ2403G. This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A).

References

- [1] Molla Imaduddin Ahmed, Brendan Spooner, John Isherwood, Mark Lane, Emma Orrock, and Ashley Dennison. A systematic review of the barriers to the implementation of artificial intelligence in healthcare. *Cureus*, 15(10), 2023. [11](#)
- [2] Fidel Alfaro-Almagro, Mark Jenkinson, Neal K Bangerter, Jesper LR Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatis N Sotiroopoulos, Saad Jbabdi, Moises Hernandez-Fernandez, Emmanuel Vallee, et al. Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *Neuroimage*, 166:400–424, 2018. [13](#)
- [3] J Andersson, S Smith, and M Jenkinson. Fnirt-fmrib’s non-linear image registration tool. *Human Brain Mapping*, 2008, 2008. [13](#)
- [4] Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga. Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3395–3404, 2020. [3](#)

- [5] 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. 2, 13
- [6] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018. 13
- [7] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019. 12
- [8] 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, 2021. 3
- [9] Cosmin I Bercea, Michael Neumayr, Daniel Rueckert, and Julia A Schnabel. Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models. *arXiv preprint arXiv:2305.19643*, 2023. 10
- [10] Cosmin I Bercea, Esther Puyol-Antón, Benedikt Wiestler, Daniel Rueckert, Julia A Schnabel, and Andrew P King. Bias in unsupervised anomaly detection in brain mri. In *Workshop on Clinical Image-Based Procedures*, pages 122–131. Springer, 2023. 3, 11, 12
- [11] Cosmin I Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A Schnabel. Diffusion models with implicit guidance for medical anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 211–220. Springer, 2024. 3, 4
- [12] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtac ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019. 3
- [13] Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A Lozano. A review on outlier/anomaly detection in time series data. *ACM computing surveys (CSUR)*, 54(3):1–33, 2021. 3
- [14] Susan Y Bookheimer, David H Salat, Melissa Terpstra, Beau M Ances, Deanna M Barch, Randy L Buckner, Gregory C Burgess, Sandra W Curtiss, Mirella Diaz-Santos, Jennifer Stine Elam, et al. The lifespan human connectome project in aging: an overview. *Neuroimage*, 185:335–348, 2019. 3, 13
- [15] Stefano Cerri, Oula Puonti, Dominik S Meier, Jens Wuerfel, Mark Mühlau, Hartwig R Siebner, and Koen Van Leemput. A contrast-adaptive method for simultaneous whole-brain and lesion segmentation in multiple sclerosis. *Neuroimage*, 225:117471, 2021. 2
- [16] Sidhant Chopra, Carrissa V Cocuzza, Connor Lawhead, Jocelyn A Ricard, Loïc Labache, Lauren M Patrick, Poornima Kumar, Arielle Rubenstein, Julia Moses, Lia Chen, et al. The trans-diagnostic connectome project: a richly phenotyped open dataset for advancing the study of brain-behavior relationships in psychiatry. *medRxiv*, 2024. 13
- [17] Fergus Davnall, Connie SP Yip, Gunnar Ljungqvist, Mariyah Selmi, Francesca Ng, Bal Sanghera, Balaji Ganeshan, Kenneth A Miles, Gary J Cook, and Vicky Goh. Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? *Insights into imaging*, 3(6):573–589, 2012. 1

- [18] Peter Dayan and Laurence F Abbott. *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT press, 2005. 1
- [19] Alessandro Delmonte, Corentin Mercier, Johan Pallud, Isabelle Bloch, and Pietro Gori. White matter multi-resolution segmentation using fuzzy set theory. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 459–462. IEEE, 2019. 12
- [20] 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*, pages 9737–9746, 2022. 3, 10, 16
- [21] Arjun D Desai, Garry E Gold, Brian A Hargreaves, and Akshay S Chaudhari. Technical considerations for semantic segmentation in mri using convolutional neural networks. *arXiv preprint arXiv:1902.01977*, 2019. 11
- [22] Alexander Frotscher, Jaivardhan Kapoor, Thomas Wolfers, and Christian F Baumgartner. Unsupervised anomaly detection using aggregated normative diffusion. *arXiv preprint arXiv:2312.01904*, 2023. 3, 4, 5, 9, 10, 14, 15
- [23] Fabio Galbusera and Andrea Cina. Image annotation and curation in radiology: an overview for machine learning practitioners. *European Radiology Experimental*, 8(1):11, 2024. 2
- [24] Michael S Gazzaniga. *The cognitive neurosciences*. MIT press, 2009. 1
- [25] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik De Leeuw, Clare M Tempany, Bram Van Ginneken, et al. Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 516–524. Springer, 2017. 2
- [26] Matthew F Glasser, Stamatios N Sotiroopoulos, J Anthony Wilson, Timothy S Coalson, Bruce Fischl, Jesper L Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R Polimeni, et al. The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124, 2013. 13
- [27] Sachin Goyal, Aditi Raghunathan, Moksh Jain, Harsha Vardhan Simhadri, and Prateek Jain. Drocc: Deep robust one-class classification. In *International conference on machine learning*, pages 3711–3721. PMLR, 2020. 5
- [28] Günther Grabner, Andrew L Janke, Marc M Budge, David Smith, Jens Pruessner, and D Louis Collins. Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2006: 9th International Conference, Copenhagen, Denmark, October 1–6, 2006. Proceedings, Part II* 9, pages 58–66. Springer, 2006. 13
- [29] Alan Hájek. The reference class problem is your problem too. *Synthese*, 156(3):563–585, 2007. 3
- [30] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019. 3
- [31] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017. 11
- [32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33*, 2020. 15

- [33] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pages 13213–13232. PMLR, 2023. [15](#), [28](#), [30](#)
- [34] Juan Eugenio Iglesias, Cheng-Yi Liu, Paul M Thompson, and Zhiowen Tu. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE transactions on medical imaging*, 30(9):1617–1634, 2011. [13](#)
- [35] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnunet: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018. [2](#)
- [36] Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, 2002. [13](#)
- [37] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012. [13](#)
- [38] Jorge Jovicich, Silvester Czanner, Xiao Han, David Salat, Andre van der Kouwe, Brian Quinn, Jenni Pacheco, Marilyn Albert, Ronald Killiany, Deborah Blacker, et al. Mri-derived measurements of human subcortical, ventricular and intracranial brain volumes: reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *Neuroimage*, 46(1):177–192, 2009. [2](#)
- [39] Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven Siegelbaum, A James Hudspeth, Sarah Mack, et al. *Principles of neural science*, volume 4. McGraw-hill New York, 2000. [1](#)
- [40] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1):195, 2019. [9](#)
- [41] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. [15](#)
- [42] Igor Koval, J-B Schiratti, Alexandre Routier, Michael Bacci, Olivier Colliot, Stéphanie Allassonnière, Stanley Durleman, and Alzheimer’s Disease Neuroimaging Initiative. Statistical learning of spatiotemporal patterns from longitudinal manifold-valued networks. In *International conference on medical image computing and computer-assisted intervention*, pages 451–459. Springer, 2017. [12](#)
- [43] Julian Krebs, Hervé Delingette, Boris Mailhé, Nicholas Ayache, and Tommaso Mansi. Learning a probabilistic model for diffeomorphic registration. *IEEE transactions on medical imaging*, 38(9):2165–2176, 2019. [12](#)
- [44] Hugo Kuijf, Matthijs Biesbroek, Jeroen de Bresser, Rutger Heinen, Christopher Chen, Wiesje van der Flier, Barkhof, Max Viergever, and Geert Jan Biessels. Data of the White Matter Hyperintensity (WMH) Segmentation Challenge, 2022. URL <https://doi.org/10.34894/AECRSD>. [13](#)
- [45] Ioannis Lagogiannis, Felix Meissen, Georgios Kaassis, and Daniel Rueckert. Unsupervised pathology detection: a deep dive into the state of the art. *IEEE transactions on medical imaging*, 43(1):241–252, 2023. [3](#), [10](#), [11](#), [31](#)
- [46] Jae-Hyeok Lee, Ji Young Yun, Allison Gregory, Penelope Hogarth, and Susan J Hayflick. Brain mri pattern recognition in neurodegeneration with brain iron accumulation. *Frontiers in Neurology*, 11:1024, 2020. [12](#)
- [47] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. [5](#)

- [48] Ziyun Liang, Xiaoqing Guo, J Alison Noble, and Konstantinos Kamnitsas. Itermask 2: Iterative unsupervised anomaly segmentation via spatial and frequency masking for brain lesions in mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 339–348. Springer, 2024. [3](#), [4](#), [10](#), [14](#)
- [49] 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. [13](#)
- [50] Yaron Lipman, Marton Havasi, Peter Holderrath, Neta Shaul, Matt Le, Brian Karrer, Ricky TQ Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024. [10](#)
- [51] Alberto Llera, Thomas Wolfers, Peter Mulders, and Christian F Beckmann. Inter-individual differences in human brain structure and morphology link to variation in demographics and behavior. *Elife*, 8:e44443, 2019. [12](#)
- [52] Imperial College London. *IXI Dataset*. <https://brain-development.org/ixi-dataset/> [Accessed: 09.04.2025]. [13](#)
- [53] Andrey Malinin, Neil Band, Alexander Ganshin, German Chesnokov, Yarin Gal, Mark J. F. Gales, Alexey Noskov, Andrey Ploskonosov, Lidiomila Prokhorenkova, Ivan Provilkov, Vatsal Raina, Vyas Raina, Denis Roginskiy, Mariya Shmatova, Panos Tigas, and Boris Yangel. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*, 2021. [13](#)
- [54] Felix Meissen, Benedikt Wiestler, Georgios Kaassis, and Daniel Rueckert. On the pitfalls of using the residual as anomaly score. In *Medical Imaging with Deep Learning*, 2021. [10](#), [12](#)
- [55] Felix Meissen, Johannes Paetzold, Georgios Kaassis, and Daniel Rueckert. Unsupervised anomaly localization with structural feature-autoencoders. In *International MICCAI Brain-lesion Workshop*, pages 14–24. Springer, 2022. [4](#), [5](#), [10](#), [16](#)
- [56] 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. [13](#)
- [57] Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiroopoulos, Jesper LR Andersson, et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523–1536, 2016. [3](#)
- [58] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes*, 15(1):210, 2022. [5](#)
- [59] Sergio Naval Marimont, Vasilis Siomos, Matthew Baugh, Christos Tzelepis, Bernhard Kainz, and Giacomo Tarroni. Ensembled cold-diffusion restorations for unsupervised anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 243–253. Springer, 2024. [3](#), [5](#), [9](#), [10](#), [14](#), [15](#)
- [60] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119, 2020. [3](#)
- [61] 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. 13
- [62] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2022. 3, 10, 16
- [63] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021. 3
- [64] Saige Rutherford, Seyed Mostafa Kia, Thomas Wolfers, Charlotte Fraza, Mariam Zabihi, Richard Dinga, Pierre Berthet, Amanda Worker, Serena Verdi, Henricus G Ruhe, et al. The normative modeling framework for computational psychiatry. *Nature protocols*, 17(7):1711–1734, 2022. 11
- [65] Saige Rutherford, Thomas Wolfers, Charlotte Fraza, Nathaniel G Harnett, Christian F Beckmann, Henricus G Ruhe, and Andre F Marquand. To which reference class do you belong? measuring racial fairness of reference classes with normative modeling. *arXiv preprint arXiv:2407.19114*, 2024. 11
- [66] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 28
- [67] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 30
- [68] International Neuroimaging Data sharing Initiative. *1000 Functional Connectomes Project*. https://fcon_1000.projects.nitrc.org/ [Accessed: 09.04.2025]. 13
- [69] Vít Škvára, Tomáš Pevný, and Václav Šmídl. Are generative deep models for novelty detection truly better? *arXiv preprint arXiv:1807.05027*, 2018. 5
- [70] Leah H Somerville, Susan Y Bookheimer, Randy L Buckner, Gregory C Burgess, Sandra W Curtiss, Mirella Dapretto, Jennifer Stine Elam, Michael S Gaffrey, Michael P Harms, Cynthia Hodge, et al. The lifespan human connectome project in development: A large-scale study of brain connectivity development in 5–21 year olds. *Neuroimage*, 183:456–468, 2018. 3, 12, 13
- [71] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 14, 28
- [72] Jason R Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A Shafto, Marie Dixon, Lorraine K Tyler, Richard N Henson, et al. The cambridge centre for ageing and neuroscience (cam-can) data repository: Structural and functional mri, meg, and cognitive data from a cross-sectional adult lifespan sample. *neuroimage*, 144: 262–269, 2017. 13
- [73] Alan J Thompson, Brenda L Banwell, Frederik Barkhof, William M Carroll, Timothy Coetzee, Giancarlo Comi, Jorge Correale, Franz Fazekas, Massimo Filippi, Mark S Freedman, et al. Diagnosis of multiple sclerosis: 2017 revisions of the mcdonald criteria. *The Lancet Neurology*, 17(2): 162–173, 2018. 2
- [74] David C Van Essen, Kamil Ugurbil, Edward Auerbach, Deanna Barch, Timothy EJ Behrens, Richard Bucholz, Acer Chang, Liyong Chen, Maurizio Corbetta, Sandra W Curtiss, et al. The human connectome project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012. 3, 13

- [75] Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1):48, 2022. [12](#)
- [76] Christian Wachinger, Anna Rieckmann, Sebastian Pölsterl, Alzheimer’s Disease Neuroimaging Initiative, et al. Detect and correct bias in multi-site neuroimaging datasets. *Medical Image Analysis*, 67:101879, 2021. [2](#), [11](#)
- [77] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The thirty-seventh asilomar conference on signals, systems & computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. [12](#)
- [78] Wentian Xu, Matthew Moffat, Thalia Seale, Ziyun Liang, Felix Wagner, Daniel Whitehouse, David Menon, Virginia Newcombe, Natalie Voets, Abhirup Banerjee, et al. Feasibility and benefits of joint learning from mri databases with different brain diseases and modalities for segmentation. *arXiv preprint arXiv:2405.18511*, 2024. [2](#)
- [79] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022. [3](#), [10](#), [16](#)
- [80] 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. [3](#)
- [81] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, 112:107706, 2021. [10](#), [14](#)
- [82] Alex P Zijdenbos, Benoit M Dawant, Richard A Margolin, and Andrew C Palmer. Morphometric analysis of white matter lesions in mr images: method and validation. *IEEE transactions on medical imaging*, 13(4):716–724, 1994. [2](#)
- [83] David Zimmerer, Fabian Isensee, Jens Petersen, Simon Kohl, and Klaus Maier-Hein. Unsupervised anomaly localization using variational auto-encoders. In *International conference on medical image computing and computer-assisted intervention*, pages 289–297. Springer, 2019. [4](#)

A Supplementary Materials

A.1 Algorithmic Performance Across Brain Lesions

Here, we report the results of the main analysis in tabular form for the BraTS20 dataset (see Tab. 3), the ATLAS dataset (see Tab. 4), the Shifts dataset (see Tab. 5), and the WMH dataset (see Tab. 6). The threshold dependent metrics [Dice] and $\text{Dice}_{\text{Estimate}}$ are calculated per individual, and the average is reported. AUPRC is calculated on the complete dataset once.

Table 3: Anomaly detection performance measured in AUPRC, [Dice] and $\text{Dice}_{\text{Estimate}}$ on the BraTS20 test dataset (332 subjects). We report the results without post-processing and when using three dimensional median filtering (MF) with a kernel size of four. In the main paper only the MF values are used in the plots.

Method	BraTS20 T1w			BraTS20 T2w		
	AUPRC	[Dice]	$\text{Dice}_{\text{Estimate}}$	AUPRC	[Dice]	$\text{Dice}_{\text{Estimate}}$
RIAD	0.100	0.148	0.144	0.148	0.203	0.203
w/ MF	0.120	0.161	0.161	0.187	0.236	0.236
IterMask	0.294	0.293	0.290	0.250	0.256	0.256
w/ MF	0.308	0.301	0.297	0.252	0.258	0.257
ANDi	0.177	0.216	0.197	0.410	0.359	0.358
w/ MF	0.264	0.265	0.232	0.518	0.453	0.425
Disyre	0.158	0.258	0.249	0.638	0.594	0.594
w/ MF	0.178	0.274	0.264	0.701	0.646	0.645
FAE	0.460	0.427	0.427	0.505	0.457	0.457
w/ MF	0.480	0.437	0.437	0.522	0.467	0.467
UniAD	0.180	0.227	0.227	0.352	0.361	0.361
w/ MF	0.198	0.239	0.239	0.386	0.381	0.381
RD	0.143	0.221	0.221	0.222	0.272	0.272
w/ MF	0.154	0.231	0.231	0.247	0.283	0.283
PatchCore	0.223	0.327	0.314	0.347	0.366	0.366
w/ MF	0.226	0.334	0.321	0.355	0.372	0.372

A.2 Algorithmic Performance Across Healthy Individuals

The false positive rates for all datasets can be found in Tab. 7 as well as the bar plot for the T2w FPRs in Fig. 7

Table 4: Anomaly detection performance measured in AUPRC, [Dice] and Dice_{Estimate} on the T1w images of the ATLAS test dataset (ATLAS-I 243 individuals, ATLAS-O 212 individuals, ATLAS-N 159 individuals). We report the results without post-processing and when using three dimensional median filtering (MF) with a kernel size of four. In the main paper only the MF values are used in the plots.

Method	ATLAS-I			ATLAS-O			ATLAS-N		
	AUPRC	[Dice]	Dice _{Estimate}	AUPRC	[Dice]	Dice _{Estimate}	AUPRC	[Dice]	Dice _{Estimate}
RIAD	0.041	0.090	0.064	0.066	0.089	0.076	0.009	0.025	0.013
w/ MF	0.050	0.129	0.081	0.105	0.126	0.097	0.015	0.045	0.017
IterMask	0.037	0.116	0.108	0.093	0.090	0.089	0.055	0.041	0.020
w/ MF	0.037	0.118	0.113	0.096	0.091	0.090	0.063	0.041	0.021
ANDi	0.222	0.188	0.183	0.119	0.143	0.138	0.033	0.074	0.042
w/ MF	0.333	0.285	0.281	0.188	0.209	0.209	0.118	0.145	0.118
Disyre	0.590	0.408	0.390	0.450	0.351	0.348	0.236	0.170	0.152
w/ MF	0.616	0.434	0.432	0.500	0.386	0.384	0.303	0.196	0.191
FAE	0.271	0.177	0.177	0.106	0.161	0.161	0.059	0.048	0.044
w/ MF	0.287	0.183	0.183	0.110	0.167	0.166	0.062	0.049	0.047
UniAD	0.214	0.173	0.162	0.115	0.162	0.162	0.043	0.040	0.030
w/ MF	0.240	0.188	0.178	0.125	0.178	0.178	0.049	0.045	0.035
RD	0.105	0.113	0.109	0.083	0.122	0.121	0.029	0.028	0.024
w/ MF	0.119	0.122	0.117	0.090	0.130	0.130	0.041	0.032	0.026
PatchCore	0.312	0.181	0.148	0.127	0.160	0.147	0.116	0.046	0.022
w/ MF	0.322	0.186	0.148	0.131	0.165	0.149	0.124	0.047	0.021

A.3 Impact of Domain Shifts and Lesion Load Variability

Here we report the p-values for the different Mann-Whitney U tests before and after multiple testing correction. They can be found in Tab. 8.

A.4 Impact of demographics on false positive rates and lesion identification.

Here we show the results for T1w images of the studied impact of inter-individual anatomical variability linked to demographic attributes such as age and sex in the HCP Aging and BraTS datasets. All additional plots can be found in Fig. 8 and all results in numeric format can be found in Tab. 10.

A.5 Influence of Network choice on Feature-based approaches

To further test the performance of the *feature-based* approaches, different pretrained networks that were trained on the classification task of ImageNet have been selected. The models and pretrained weights from torchvision have been used for all experiments. We observed that version one of the set of weights

Table 5: Anomaly detection performance measured in AUPRC, [Dice] and Dice_{Estimate} on the Shifts test dataset (68 individuals). We report the results without post-processing and when using three dimensional median filtering (MF) with a kernel size of four. In the main paper only the MF values are used in the plots.

Method	Shifts T1w			Shifts T2w		
	AUPRC	[Dice]	Dice _{Estimate}	AUPRC	[Dice]	Dice _{Estimate}
RIAD	0.021	0.050	0.038	0.020	0.051	0.040
w/ MF	0.024	0.062	0.045	0.021	0.064	0.050
IterMask	0.016	0.044	0.034	0.010	0.020	0.015
w/ MF	0.017	0.045	0.031	0.010	0.021	0.015
ANDi	0.072	0.123	0.123	0.054	0.099	0.099
w/ MF	0.110	0.143	0.125	0.061	0.128	0.104
Disyre	0.148	0.151	0.148	0.235	0.208	0.197
w/ MF	0.166	0.156	0.139	0.234	0.205	0.172
FAE	0.022	0.032	0.017	0.043	0.049	0.032
w/ MF	0.022	0.032	0.015	0.044	0.048	0.031
UniAD	0.030	0.045	0.045	0.04	0.059	0.053
w/ MF	0.033	0.048	0.048	0.046	0.063	0.055
RD	0.029	0.028	0.024	0.032	0.043	0.035
w/ MF	0.041	0.032	0.026	0.034	0.045	0.035
PatchCore	0.043	0.057	0.026	0.045	0.056	0.017
w/ MF	0.046	0.058	0.025	0.049	0.058	0.015

performed better for all anomaly detection tasks tested and that ImageNet specific normalization for mean and standard deviation on the images yielded slightly reduced performance. In Fig. 9 we additionally show FAE and PatchCore with features extracted from EfficientNet-B4 and UniAD when using features from WideResNet50. Furthermore, we updated the network of the FAE, as it was the least sophisticated network used in the general analysis. We introduced residual connections and a second convolution at each stage to increase the depth of the originally shallow network. This configuration uses the features of the pretrained EfficientNet-B4, and we call it FAE-v2. The specific combination of the structural similarity index (SSIM) with deep features used by the FAE showed remarkable consistency throughout the different networks employed. The FAE-v2 showed no improvements over the FAE. We argue that this observation points to fundamental problems in the design of UAD methods. A better performance on the pretext task usually does not correlate with downstream detection performance. This behavior is observable for two of the reconstruction-based approaches and for all feature-based approaches in the conducted analysis. In the main paper, we approach this problem by using the checkpoints that achieved maximum performance on the validation dataset to determine when the method is most valuable for the downstream task.

Table 6: Anomaly detection performance measured in AUPRC, [Dice] and Dice_{Estimate} on the T1w images of the WMH test dataset (54 individuals). We report the results without post-processing and when using three dimensional median filtering (MF) with a kernel size of four. In the main paper only the MF values are used in the plots.

Method	WMH		
	AUPRC	[Dice]	Dice _{Estimate}
RIAD	0.019	0.037	0.035
w/ MF	0.023	0.044	0.042
IterMask	0.019	0.037	0.036
w/ MF	0.021	0.037	0.036
ANDi	0.028	0.053	0.051
w/ MF	0.031	0.062	0.052
Disyre	0.085	0.119	0.112
w/ MF	0.098	0.130	0.109
FAE	0.019	0.032	0.018
w/ MF	0.019	0.032	0.016
UniAD	0.027	0.042	0.042
w/ MF	0.028	0.044	0.044
RD	0.022	0.038	0.028
w/ MF	0.023	0.039	0.027
PatchCore	0.044	0.072	0.050
w/ MF	0.045	0.074	0.049

A.6 Disyre Ablation Study

In order to understand the outstanding performance of Disyre in our benchmark, we conducted a small ablation study on the validation dataset. This analysis is similar to the scaling analysis shown in the main part of the manuscript. We decided to remove components of the Disyre method that could be used for the majority of the other methods. These include the specific preprocessing used and the patch-based paradigm. The results can be seen in Fig. 10. We observed that the patch-based paradigm could be removed without any performance loss; instead, we observed a small increase in performance. In contrast, the specific preprocessing used seems to be an integral part of the Disyre model and its ability to achieve state-of-the-art performance. The preprocessing includes an elastic transformation, various intensity transformations, and a mirror transformation. We did not study the effect of this preprocessing on the other methods and want to note that an interaction between the synthetic anomaly generation pipeline and this preprocessing is possible, i.e., only in combination is a robust way of generating synthetic anomalies possible.

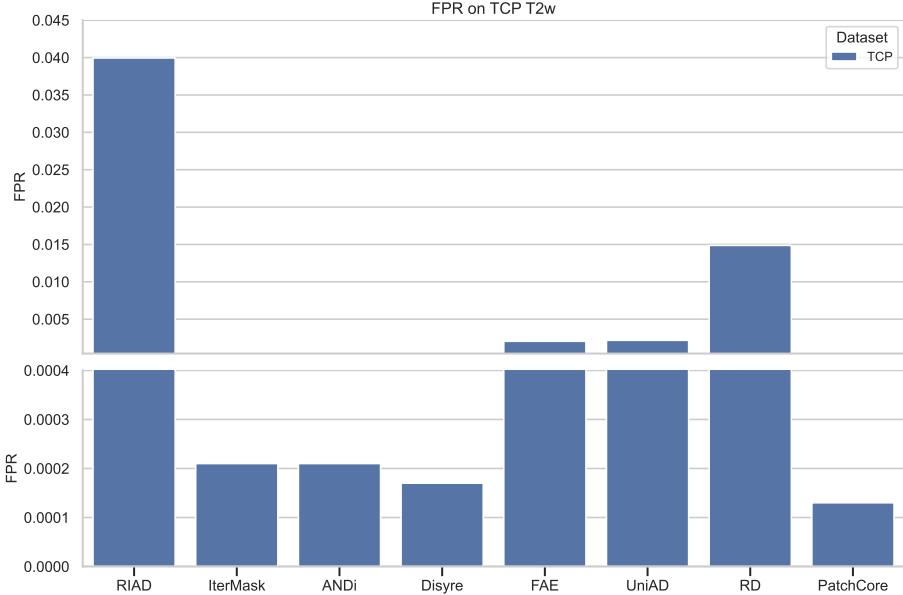


Figure 7: **False positive rates on the T2w healthy portion of the TCP datasets for all tested algorithms based on the *estimated* threshold.**

A.7 Model and Training Details

Here, we report the architectures used and method-related details. In general, we aimed to reduce heterogeneity between the methods regarding preprocessing and architectural choices while maximizing individual performances. Nevertheless, a risk remains that the methods have not achieved optimal performance due to suboptimal choices of hyperparameters or misalignment between performance on the validation dataset and the test sets. All *reconstruction-based* methods that originally used a U-Net-like architecture were equipped with the same U-Net, which is a slightly modified version of the DDPM++ version used in the diffusion model literature⁷¹. We opted for this choice because it ensures that older methods start on equal footing with the newly published methods. The DDPM++ model integrates the BigGAN residual blocks and scales the residual connections by $\frac{1}{\sqrt{2}}$. Note that we usually omit the scaling for all non-diffusion model methods. Additionally, we introduced a modification known as efficient U-Net, which swaps the order of the downsampling operation and the first convolution of each block⁶⁶, use the dropout modification proposed in³³, employ the memory efficient attention mechanism provided by PyTorch, and zero-initialize the last convolution in each block.

A.7.1 RIAD

In general, we build our RIAD implementation on the publicly available third-party code provided by <https://github.com/plutoyuxie/Reconstruction-by-inpainting-for-visual-anomaly-detection>. We removed the median filtering in the gradient magnitude similarity, as we could not find this calculation step in the original paper. Additionally, we added a small number to the square root calculation used in the

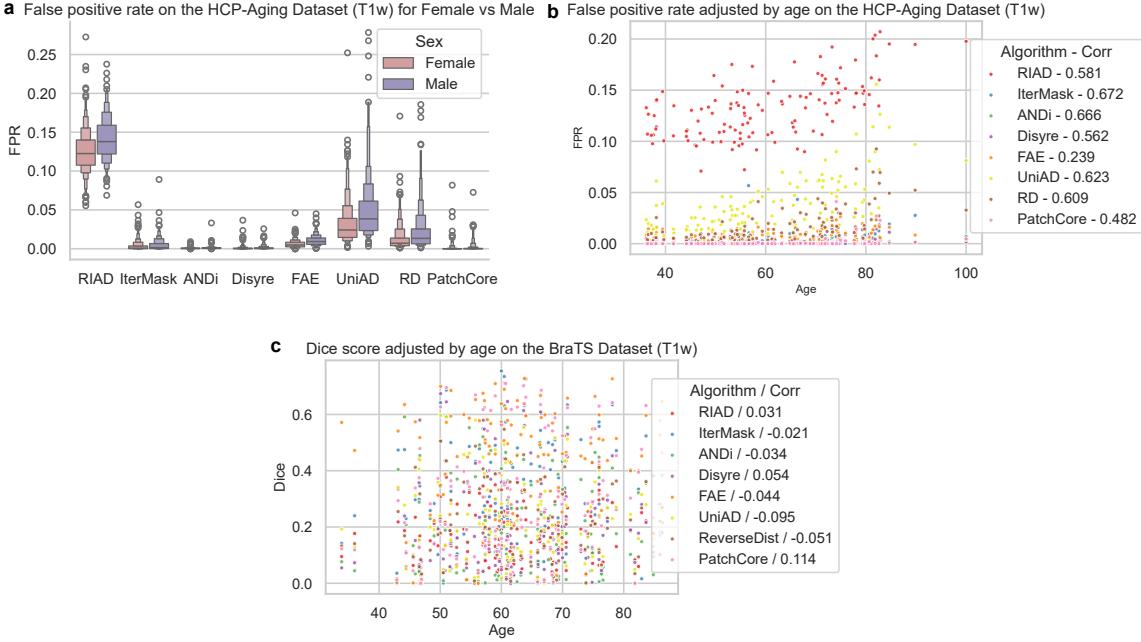


Figure 8: Impact of demographics on false positive rates and lesion identification for T1w images. **a)** The false positive rate on the HCP Aging dataset for the female and male group. All groups show significant differences based on the Mann-Whitney U test. Significance level $\alpha = 0.05$ was used for all tests. **b)** The false positive rate on the HCP Aging dataset adjusted by age. The Spearman rank correlation test displayed that all algorithms show significant positive correlation indicating that older subjects are more likely to have an anomaly assigned to a voxel that is normal. **c)** The Dice score on the BraTS dataset adjusted by age.

edge filter to reduce the numerical problems encountered during optimization. We still experienced unstable optimization for this method throughout all the experiments. Additionally, we experimented with the masking hyperparameter k to ensure coverage of large brain lesions while proportionally increasing the hyperparameter n , thereby maintaining sufficient contextual information for accurate reconstruction. The original implementation used $n = 3$ and $k = [2, 4, 8, 16]$. All hyperparameters can be found in Tab. 11.

A.7.2 IterMask

The IterMask implementation is built on the officially available implementation (<https://github.com/ZiyunLiang/IterMask2>). For this method, we scaled the image slices that are in the range $[0,1]$ with $\text{slice} \cdot 6 - 3$ to bring the slices to $[-3, 3]$, resulting in a maximum intensity number of 3 in each brain. Note that min/max normalization has been applied to the complete brain before this step. The Gaussian mask generation is performed on the complete 240×240 slice and then center cropped along with the rest of the image slice to 224×224 . For mask generation and frequency masking, we keep all hyperparameters of the official implementation. All hyperparameters for both models are found in Tab. 12.

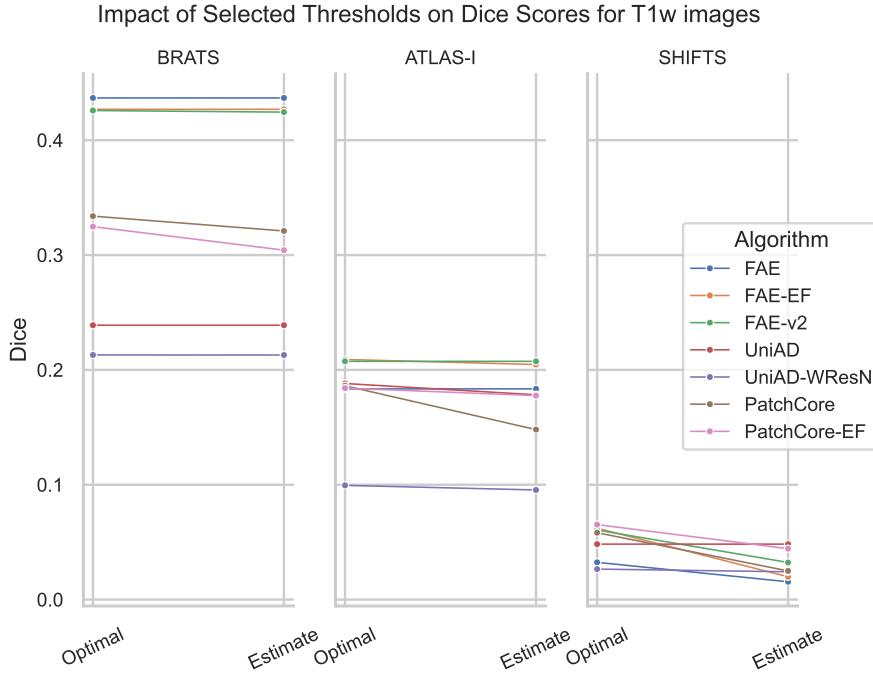


Figure 9: **Impact of the pretrained network on state-of-the-art *feature-based* unsupervised anomaly detection.** The performance of the algorithms on T1w images were reported using two different thresholds, the *optimal* threshold, defined as the maximum possible Dice score optimized in the test set, thus potentially susceptible to bias but standard in the field. Second, the *estimated* threshold optimized on the validation data set, then fixed and performance for that threshold reported on the untouched test set, thus unbiased but not the standard in the field. Here we tested multiple *feature-based* approaches while altering the pretrained network.

A.7.3 ANDi

ANDi was developed by the first author of this paper, and this newly provided implementation can be seen as the up-to-date variant of the method. We mainly followed the diffusion model implementation of^{33;67} and used a Variational Diffusion Model (VDM) with a cosine schedule that is adjusted for the image size used, as in³³. We used 56x56 as the base size for the shifting operation. Additionally, we use velocity prediction during training and learn to minimize the distance between the ground truth noise and the predicted noise, which is calculated from the velocity. During inference, we calculate the difference between the ground truth latent and the predicted latent, as in the original ANDi implementation. We aggregate the individual deviations using the geometric mean and change the Pyramid noise to standard Gaussian noise. The new noise schedule requires different values for the parameters T_l, T_u and we have observed improvements in anomaly detection performance upon introducing this change. For the Pyramid noise, we opt for $c = 0.9$ and image slices are scaled to be in the range of [-1, 1] before using the network or adding the noise. All hyperparameters can be found in Tab. 13.

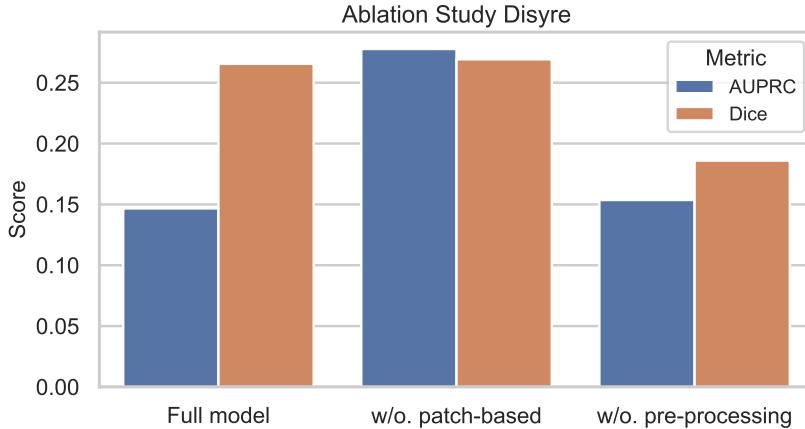


Figure 10: **Disyre Ablation Study.** Three different variants of the Disyre method have been evaluated on the validation dataset. First the full model, a model without the patch-based paradigm that takes as input the full slice and a model that omits the Disyre specific preprocessing. We observed that the specific preprocessing used is important for the Disyre model to achieve its optimum performance.

A.7.4 Disyre

We built our Disyre implementation on the official implementation (<https://github.com/snavalm/disyre>) and used Disyre v2 for all experiments. We downloaded the shapes provided on GitHub (needed for the DAG) and kept all hyperparameters of the official implementation, besides the tissue classes estimated with k-means and the architecture of the network. For the tissue classes, we run k-means on our training dataset for T1w and T2w images separately and use the respective clusters for the different training runs. As mentioned in Sec. A.6, we maintain the specific preprocessing pipeline of Disyre as well as the patch-based paradigm for our implementation and provide results for ablations regarding these options. After the DAG pipeline, we bring the slices to $[-1, 1]$ to follow the typical diffusion model literature. All hyperparameters can be found in Tab. 14.

A.7.5 FAE

For FAE, we used the implementation provided by the UPD study⁴⁵ (https://github.com/iolag/UPD_study/), as the authors of this benchmark are also the authors of the original article. For training, we use the SSIM with a window size of five, whereas for testing, we use a window size of 11. We experienced better results with this setup on the larger images (feature maps) and the architecture of the network is adapted to work with this larger resolution. We observed no difference when using the ImageNet specific feature normalization. All hyperparameters can be found in Tab. 15.

A.7.6 UniAD

We built our implementation of UniAD on the official code available at <https://github.com/zhiyuanyou/UniAD>. We have tried different variants of UniAD by altering the feature size and masking strategy used in the modified attention mechanism but ultimately found the original hyperparameters to work best. In our

experiments, it was crucial not to take many gradient update steps and not to use feature normalization. The hyperparameters can be found in Tab. 16.

A.7.7 Reverse Distillation

The RD implementation is built on the official code provided by <https://github.com/hq-deng/RD4AD>. We adapted the code to work with the image resolution used in the experiments and noticed that feature normalization and training for too long hurt the method’s performance drastically. Hyperparameters can be found in Tab. 17.

A.7.8 PatchCore

The implementation of PatchCore follows the official code that can be found at <https://github.com/amazon-science/patchcore-inspection>. PatchCore is the only method that does not require any training and uses a memory bank instead. We randomly sampled 92 volumes from our training dataset to build the memory bank. Note that increasing the number of volumes would drastically increase the time required to build the memory bank. When using the 92 volumes and the greedy coresset subsampling to reduce the size to 1%, the time to build the bank is roughly three days. We experimented with different patch sizes, strides, and memory bank sizes but found the setup working for MVTec anomaly localization to be well suited for our framework as well. Slight improvements were observed when the number of neighbors to search for in the memory bank was set to one. Hyperparameters can be found in Tab. 18.

Table 7: False positive rates on all datasets (FCON1T 139 individuals, FCON3T 105 individuals, TCP 92 individuals). We report the results without post-processing and when using three dimensional median filtering (MF) with a kernel size of four. In the main paper only the MF values are used in the plots.

Method	Datasets				
	FCON1T	FCON3T	TCP T1w	TCP T2w	
RIAD	2.00×10^{-1}	1.74×10^{-1}	1.46×10^{-1}	8.61×10^{-2}	
w/ MF	1.23×10^{-1}	1.06×10^{-1}	8.30×10^{-2}	4.00×10^{-2}	
IterMask	1.67×10^{-1}	1.50×10^{-1}	2.15×10^{-2}	5.50×10^{-4}	
w/ MF	1.60×10^{-1}	1.42×10^{-1}	1.71×10^{-2}	2.10×10^{-4}	
ANDi	3.71×10^{-2}	2.80×10^{-2}	1.18×10^{-2}	8.30×10^{-3}	
w/ MF	4.82×10^{-3}	5.83×10^{-3}	3.79×10^{-3}	2.10×10^{-4}	
Disyre	2.56×10^{-3}	5.00×10^{-3}	1.69×10^{-3}	1.08×10^{-3}	
w/ MF	8.20×10^{-4}	1.99×10^{-3}	4.00×10^{-4}	1.70×10^{-4}	
FAE	1.77×10^{-2}	1.49×10^{-2}	3.01×10^{-3}	2.67×10^{-3}	
w/ MF	1.44×10^{-2}	1.23×10^{-2}	2.23×10^{-3}	2.04×10^{-3}	
UniAD	1.65×10^{-2}	1.51×10^{-2}	1.04×10^{-2}	4.41×10^{-3}	
w/ MF	1.12×10^{-2}	1.01×10^{-2}	6.64×10^{-3}	2.17×10^{-3}	
RD	9.12×10^{-2}	3.91×10^{-2}	1.77×10^{-2}	1.93×10^{-2}	
w/ MF	7.95×10^{-2}	3.16×10^{-2}	1.29×10^{-2}	1.49×10^{-2}	
PatchCore	2.09×10^{-2}	5.56×10^{-3}	4.20×10^{-4}	2.70×10^{-4}	
w/ MF	1.84×10^{-2}	4.69×10^{-3}	3.20×10^{-4}	1.30×10^{-4}	

Table 8: P-value, and corrected p-value using the Benjamini–Hochberg method for the Mann-Whitney U tests between Dice scores for ID and OOD pairs. We constructed eight datasets out of the ATLAS-I and ATLAS-O datasets corresponding to four matched ID–OOD pairs, stratified by lesion load percentiles derived from the full ATLAS cohort. We group volumes above the 75th percentile (Top), above the 50th percentile to 75th (Upper), above the 25th percentile to 50th percentile (Middle), and below the 25th percentile (Lower). The reported values are derived from median filtered anomaly maps and the *estimated* threshold.

Algorithm	Top	Upper	Middle	Lower
<i>standard p-value</i>	5.12×10^{-1}	9.42×10^{-2}	3.14×10^{-1}	2.49×10^{-1}
	5.89×10^{-3}	4.81×10^{-2}	8.96×10^{-2}	1.77×10^{-2}
	1.88×10^{-8}	1.44×10^{-1}	1.21×10^{-1}	1.60×10^{-1}
	4.48×10^{-9}	2.35×10^{-2}	5.28×10^{-1}	7.25×10^{-2}
	5.59×10^{-2}	1.35×10^{-2}	1.72×10^{-2}	7.13×10^{-1}
	1.02×10^{-1}	3.06×10^{-1}	2.16×10^{-1}	4.86×10^{-1}
	3.71×10^{-1}	1.35×10^{-1}	5.59×10^{-2}	4.98×10^{-1}
	8.96×10^{-1}	5.06×10^{-2}	6.07×10^{-1}	4.91×10^{-1}
<i>corrected p-value</i>	5.12×10^{-1}	3.77×10^{-1}	4.19×10^{-1}	4.19×10^{-1}
	2.35×10^{-2}	6.42×10^{-2}	8.96×10^{-2}	3.54×10^{-2}
	7.51×10^{-8}	1.60×10^{-1}	1.60×10^{-1}	1.60×10^{-1}
	1.79×10^{-8}	4.70×10^{-2}	5.28×10^{-1}	9.67×10^{-2}
	7.45×10^{-2}	3.45×10^{-2}	3.45×10^{-2}	7.13×10^{-1}
	4.08×10^{-1}	4.08×10^{-1}	4.08×10^{-1}	4.86×10^{-1}
	4.95×10^{-1}	2.69×10^{-1}	2.24×10^{-1}	4.98×10^{-1}
	8.96×10^{-1}	2.03×10^{-1}	8.09×10^{-1}	8.09×10^{-1}

Table 9: P-values for the independent variables and the interaction term as well as the degrees of freedom (site, size, interaction) used for the Chi-Square approximation for the Scheirer–Ray–Hare tests. The Dice scores for each method were selected as the dependent variable and the lesion load categories and the ID-OOD dummy variable as the independent variables. The reported values are derived from median filtered anomaly maps and the *estimated* threshold.

Algorithm	p-value Site	p-value Lesion Size	p-value Interaction	Degrees of Freedom
RIAD	7.10×10^{-1}	0.0×10^0	9.68×10^{-1}	1, 3, 3
IterMask	5.99×10^{-1}	0.0×10^0	3.58×10^{-1}	1, 3, 3
ANDi	4.68×10^{-5}	0.0×10^0	2.05×10^{-1}	1, 3, 3
Disyre	9.93×10^{-4}	0.0×10^0	9.99×10^{-3}	1, 3, 3
FAE	5.36×10^{-2}	0.0×10^0	7.55×10^{-1}	1, 3, 3
UniAD	1.77×10^{-1}	0.0×10^0	9.78×10^{-1}	1, 3, 3
RD	3.94×10^{-1}	0.0×10^0	8.95×10^{-1}	1, 3, 3
PatchCore	5.15×10^{-1}	0.0×10^0	5.91×10^{-1}	1, 3, 3

Table 10: P-value, mean, standard deviation and Cohen’s d of FPR values calculated on HCP Aging for all algorithms. The reported values are derived from median filtered anomaly maps.

Algorithm	p-value	Mean male	Mean female	Std male	Std female	Cohen’s d
<i>T_{1w} images</i>	RIAD	5.88×10^{-15}	1.41×10^{-1}	1.25×10^{-1}	2.78×10^{-2}	2.79×10^{-2}
	IterMask	1.99×10^{-4}	4.87×10^{-3}	3.34×10^{-3}	8.40×10^{-3}	6.43×10^{-3}
	ANDi	1.10×10^{-5}	1.15×10^{-3}	7.66×10^{-4}	2.26×10^{-3}	1.08×10^{-3}
	Disyre	1.84×10^{-8}	1.38×10^{-3}	1.01×10^{-3}	2.76×10^{-3}	2.84×10^{-3}
	FAE	1.19×10^{-25}	1.05×10^{-2}	5.97×10^{-3}	6.72×10^{-3}	5.15×10^{-3}
	UniAD	4.19×10^{-17}	4.95×10^{-2}	3.16×10^{-2}	4.02×10^{-2}	2.69×10^{-2}
	RD	2.53×10^{-14}	2.20×10^{-2}	1.24×10^{-2}	2.66×10^{-2}	1.57×10^{-2}
	PatchCore	1.50×10^{-4}	1.66×10^{-3}	1.10×10^{-3}	6.10×10^{-3}	5.65×10^{-3}
<i>T_{2w} images</i>	RIAD	4.15×10^{-7}	6.70×10^{-2}	5.76×10^{-2}	2.81×10^{-2}	2.66×10^{-2}
	IterMask	7.91×10^{-3}	1.53×10^{-3}	9.98×10^{-4}	3.63×10^{-3}	2.24×10^{-3}
	ANDi	2.49×10^{-3}	6.85×10^{-4}	5.80×10^{-4}	1.36×10^{-3}	1.06×10^{-3}
	Disyre	4.66×10^{-6}	5.35×10^{-4}	3.40×10^{-4}	1.49×10^{-3}	7.66×10^{-4}
	FAE	1.02×10^{-10}	6.37×10^{-3}	4.41×10^{-3}	8.52×10^{-3}	7.06×10^{-3}
	UniAD	1.29×10^{-10}	9.57×10^{-3}	5.72×10^{-3}	1.33×10^{-2}	7.69×10^{-3}
	RD	5.79×10^{-8}	7.80×10^{-3}	4.82×10^{-3}	1.17×10^{-2}	7.54×10^{-2}
	PatchCore	1.54×10^{-4}	1.61×10^{-3}	8.61×10^{-4}	5.74×10^{-3}	3.84×10^{-3}

Table 11: RIAD Hyperparameters

Hyperparameter	Value
α	1
β	1
γ	1
k	[8,16,28,32]
n	5
Update Steps	100,000
Batch size	32
Optimizer	AdamW
β_1, β_2	0.9, 0.999
Weight Decay	0.01
lr	$5e - 6$
Image Size	224x224
Number of Input Channels	1
Model Base Dimension (Channels)	32
Channel multiplier per Resolution	(1, 1, 1, 2, 3, 4)
Number of Blocks per Resolution	(1, 1, 1, 2, 3, 2)
Nonlinearity	Swish
Normalization	GroupNorm
Attention Resolution	(14, 7)
Attention Embedding Dimension	32
Dropout	0.1
Dropout Start Resolution	14
Rescale residual connections	False
Final checkpoint (number of update steps)	100,000

Table 12: IterMask Hyperparameters

Hyperparameter	First Model	Second Model
Update Steps	100,000	100,000
Batch size	128	128
Optimizer	AdamW	AdamW
β_1, β_2	0.9, 0.999	0.9, 0.999
Weight Decay	0.01	0.01
lr	$1e - 4$	$1e - 4$
EMA rate	0.9999	0.9999
Image Size	224x224	224x224
Number of Input Channels	1	2
Model Base Dimension (Channels)	32	32
Channel multiplier per Resolution	(1, 1, 1, 2, 3, 4)	(1, 1, 1, 2, 3, 4)
Number of Blocks per Resolution	(1, 1, 1, 2, 4, 2)	(1, 1, 1, 2, 4, 2)
Nonlinearity	Swish	Swish
Normalization	BatchNorm	BatchNorm
Attention Resolution	(14, 7)	(14, 7)
Attention Embedding Dimension	32	32
Dropout	0.1	0.1
Dropout Start Resolution	14	14
Rescale residual connections	False	False
Final checkpoint (number of update steps)	100,000	100,000

Table 13: ANDi Hyperparameters

Hyperparameter	Value
Number of Latent Variables	1000
Log SNR Minimum	-15
Log SNR Maximum	15
V-Prediction	True
Shift Schedule	True
Time Embedding Type	Positional
Pyramid Discount	0.9
T_l	25
T_u	125
Update Steps	100,000
Batch size	128
Optimizer	AdamW
β_1, β_2	0.9, 0.999
Weight Decay	0.01
lr	$1e - 4$
EMA rate	0.9999
Image Size	224x224
Number of Input Channels	1
Model Base Dimension (Channels)	32
Channel multiplier per Resolution	(1, 1, 1, 2, 3, 4)
Number of Blocks per Resolution	(1, 1, 1, 2, 4, 2)
Nonlinearity	Swish
Normalization	GroupNorm
Attention Resolution	(14, 7)
Attention Embedding Dimension	32
Dropout	0.1
Dropout Start Resolution	14
Rescale residual connections	True
Final checkpoint (number of update steps)	30,000

Table 14: Disyre Hyperparameters

Hyperparameter	Value
Number of Latent Variables	100
Beta Minimum	0.1
Beta Maximum	20
Time Embedding Type	Positional
Anomaly Type	DAG
Anomaly Patch Size	64x64
No Anomaly in Background	True
T1w clusters	[0.6079509, 0.07953554, 0.4649098, 0.78062236, 0.31114596]
T2w clusters	[0.04920235, 0.5074527, 0.2573119, 0.7149557, 0.37794548]
Update Steps	100,000
Batch size	64
Optimizer	AdamW
β_1, β_2	0.9, 0.999
Weight Decay	0.01
lr	$1e - 4$
EMA rate	0.9999
Image Size (Patch Size)	128x128
Number of Input Channels	1
Model Base Dimension (Channels)	64
Channel multiplier per Resolution	(1, 2, 2, 4, 4)
Number of Blocks per Resolution	(1, 1, 2, 4, 2)
Nonlinearity	Swish
Normalization	GroupNorm
Attention Resolution	(16, 8)
Attention Embedding Dimension	64
Dropout	0.1
Dropout Start Resolution	8
Rescale residual connections	True
Final checkpoint (number of update steps)	100,000

Table 15: FAE Hyperparameters

Hyperparameter	Value
Update Steps	40,000
Batch size	128
Optimizer	AdamW
β_1, β_2	0.9, 0.999
Weight Decay	0.01
lr	$1e - 4$
Image Size	224x224
Model Channels	[100, 150, 200, 300]
Dropout	0.1
Pretrained Network	ResNet18
Feature Map Size	56x56
Extracted Layers	['maxpool', 'layer1', 'layer2']
Final checkpoint T1w (number of update steps)	20,000
Final checkpoint T2w (number of update steps)	40,000

Table 16: UniAD Hyperparameters

Hyperparameter	Value
Update Steps	100,000
Batch size	128
Optimizer	AdamW
β_1, β_2	0.9, 0.999
Weight Decay	0.01
lr	$1e - 4$
lr scheduler	Linear
Clip Gradient Norm	0.1
Image Size	224x224
Pretrained Network	EfficientNet-B4
Feature Map Size	14x14
Extracted Layers	['features.1', 'features.2', 'features.3', 'features.4']
Position Embedding	Learned
Model Hidden Dimension	256
Model Number of Heads	8
Number Encoder Layers	4
Number Decoder Layers	4
Feedforward Dimension	1024
Dropout	0.1
Nonlinearity	ReLU
Feature Jitter	True
Feature Jitter Scale	20.0
Jitter Probability	1.0
Neighbor Size	[7, 7]
Neighbor Mask	[True, True, True]
Final checkpoint (number of update steps)	25,000

Table 17: RD Hyperparameters

Hyperparameter	Value
Update Steps	25,000
Batch size	128
Optimizer	AdamW
β_1, β_2	0.9, 0.999
Weight Decay	0.01
lr	$1e - 4$
Image Size	224x224
Pretrained Network	ResNet18
Extracted Layers	['layer1', 'layer2', 'layer3']
Final checkpoint T1w (number of update steps)	15,000
Final checkpoint T2w (number of update steps)	20,000

Table 18: PatchCore Hyperparameters

Hyperparameter	Value
Image Size	224x224
Pretrained Network	WideResNet50
Extracted Layers	['layer2', 'layer3']
Number of Neighbors	1
Patch Size	5
Patch Stride	1
Target Embedding Dimension	1024
Sample Size	0.01