Journal of Machine Learning for Biomedical Imaging 2024:020 vol. 2, pp. 2006–2052
Special issue: UNSURE2023 special issue
Guest editors: Christian Baumgartner, Adrian Dalca, Raghav Mehta, Chen Qin, Carole Sudre, William (Sandy) Wells

Submitted 12/2023
Published 10/2024

Dimensionality Reduction and Nearest Neighbors for Improving Out-of-Distribution Detection in Medical Image Segmentation

McKell Woodland https://orcid.org/0000-0003-0995-3695 mewoodland@mdanderson.org
The University of Texas MD Anderson Cancer Center, Houston, TX, USA
Rice University, Houston, TX, USA
Nihil Patel https://orcid.org/0000-0002-5382-1983
Austin Castelo https://orcid.org/0009-0009-3015-3570
Mais Al Taie https://orcid.org/0009-0000-9298-6644
Mohamed Eltaher https://orcid.org/0009-0008-2078-4718
Joshua P. Yung https://orcid.org/0000-0003-3752-5997
Tucker J. Netherton https://orcid.org/0000-0003-1583-7121
Tiffany L. Calderone https://orcid.org/0000-0003-4404-5342
Jessica I. Sanchez https://orcid.org/0009-0006-1042-5661
The University of Texas MD Anderson Cancer Center, Houston, TX, USA
Darrel W. Cleere
Ahmed Elsaiey
Nakul Gupta https://orcid.org/0000-0002-2413-2193
David Victor https://orcid.org/0000-0003-1414-3128
Houston Methodist Hospital, Houston, TX, USA
Laura Beretta https://orcid.org/0000-0002-2054-684X
The University of Texas MD Anderson Cancer Center, Houston, TX, USA
Ankit B. Patel https://orcid.org/0000-0001-9678-496X
Baylor College of Medicine, Houston, TX, USA
Rice University, Houston, TX, USA
Kristy K. Brock https://orcid.org/0000-0001-9364-5040 kkbrock@mdanderson.org
The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Abstract

Clinically deployed deep learning-based segmentation models are known to fail on data outside of their training distributions. While clinicians review the segmentations, these models tend to perform well in most instances, which could exacerbate automation bias. Therefore, detecting out-of-distribution images at inference is critical to warn the clinicians that the model likely failed. This work applied the Mahalanobis distance (MD) post hoc to the bottleneck features of four Swin UNETR and nnU-net models that segmented the liver on T1-weighted magnetic resonance imaging and computed tomography. By reducing the dimensions of the bottleneck features with either principal component analysis or uniform manifold approximation and projection, images the models failed on were detected with high performance and minimal computational load. In addition, this work explored a non-parametric alternative to the MD, a k-th nearest neighbors distance (KNN). KNN drastically improved scalability and performance over MD when both were applied to raw and average-pooled bottleneck features. Our code is available at `https://github.com/mckellwoodland/dimen_reduce_mahal`.

Keywords: Out-of-distribution detection, Mahalanobis distance, Nearest Neighbors, Principal component analysis, Uniform manifold approximation and projection

https://doi.org/10.59275/j.melba.2024-g93a

1. Introduction

Liver malignancy is one of the leading causes of cancer death worldwide (Ferlay et al., 2021), with mortality rates increasing more rapidly than all other cancers within the United States (Ryerson et al., 2016). Radiotherapy is a non-invasive treatment for advanced liver cancer that leverages ionizing radiation to treat tumors (Chen et al., 2020). Precise delineation of treatment targets and surrounding anatomical structures is critical to the success of radiotherapy. Manual segmentation of these structures is time-intensive (Multi-Institutional, 2011), leading to delays that have been correlated with lower survival rates (Chen et al., 2008) and incompatibility with techniques that require frequent imaging to account for anatomical changes (Sheng, 2020), such as magnetic resonance imaging (MRI)-guided adaptive radiotherapy (Otazo et al., 2021). In addition, manual segmentation is subject to human variability and inconsistencies (Nelms et al., 2012), which can lead to a lower quality of radiotherapy (Saarnak et al., 2000). These limitations have prompted expansive research into automated segmentation methods.

Deep learning (DL) algorithms constitute the current state-of-the-art for medical imaging segmentation, with research spanning many anatomical regions and imaging modalities (Cardenas et al., 2019). However, DL models struggle to generalize to information that was not present while the model was being trained (Zech et al., 2018). This problem is exacerbated in the medical field, where collecting large-scale, annotated, and diverse training datasets is challenging due to the cost of labeling, rare cases, and patient privacy. Even models with high performance during external validation may fail when presented with novel information after clinical deployment. This can be demonstrated by the work of Anderson et al. (2021). On test data, 96% of their DL-based liver segmentations were deemed clinically acceptable, with most of their segmentations being preferred over manual segmentations. The two images the model performed poorly on contained information not present during training – namely, the presence of ascites and a stent.

Automated segmentations are typically manually evaluated and corrected, if need be, by a clinician before they are used in patient treatment. The main concern with human evaluation is automation bias, where physicians may become too reliant on model output. To protect against automation bias, it is critical to warn clinicians of potential segmentation model failure. Identifying model inputs that will lead to poor model performance is referred to as out-of-distribution (OOD) detection (Yang et al., 2024). This study focuses on post-hoc OOD detection or methods that can be applied after model training in order to develop warning systems for models already in clinical deployment.

Mahalanobis distance (MD) is a commonly used post-hoc OOD detection method that computes the distance between a test image and a Gaussian distribution fitted to training images (Lee et al., 2018). Given the inherent high dimensionality of images, the distance is typically applied to features extracted from the network being analyzed. MD has achieved state-of-the-art performance in natural imaging when applied directly to classifier features (Fort et al., 2021). However, features from medical imaging segmentation models are an order of magnitude larger than classifier features, necessitating further dimensionality reduction to ensure computational feasibility. While average pooling has been used conventionally to reduce feature dimensionality (Lee et al., 2018; González et al., 2021), no prior studies have examined the best way to prepare features for the MD calculation. Other open areas of research in regards to the application of MD to medical imaging segmentation models include the validity of the Gaussian assumption, which features from the model should be utilized (González et al., 2022), how to combine features best if multiple features are utilized, and how to utilize MD with multi-class segmentation networks.

We aim to improve the performance and scalability of feature-based OOD detection in medical imaging segmentation. Our main contributions are two-fold. First, we propose using principal component analysis (PCA) and uniform manifold approximation and projection (UMAP) to prepare features for the MD calculation. We demonstrate that these methods outperform average pooling across four liver segmentation models. Second, we propose using a k-th nearest neighbor (KNN) distance (Sun et al., 2021) as a distribution-agnostic replacement for MD for medical imaging segmentation models. Our results show a drastic improvement of KNN over MD on raw and average pooled features, questioning the validity of the Gaussian assumption for segmentation model features.

This work was first published in Lecture Notes of Computer Science volume 14291, pages 147–156 by Springer Nature (Woodland et al., 2023). It was extended to include validation of the dimensionality reduction techniques for three additional liver segmentation models (including extensions to computed tomography and the nnU-net architecture). Furthermore, the extension includes a novel analysis of the KNN distance as a replacement for MD. Finally, the extension provides greater context into how MD and KNN fit into the larger OOD detection field by comparing their performance to standard methods.

2. Related Works

Traditional OOD detection aims to identify and reject model input whose true label deviates semantically from the label distribution observed during the model's training phase (Yang et al., 2024). Our work follows an alternative definition of OOD detection common in many safety-critical applications: identifying and rejecting model input that falls outside the model's generalization capacity (Pleiss et al., 2019; Yang et al., 2024). Distribution in this context refers to a theoretical statistical distribution where data drawn from it is within the scope of the model under consideration. A comprehensive review of OOD detection approaches was recently compiled by Yang et al. (2024). Our review focuses on feature- and output-based OOD detection methods that can be applied post hoc to segmentation models and training-based uncertainty estimation approaches.

Lee et al. (2018) utilized density estimation for OOD detection by calculating the MD between test features and class-conditional Gaussian distributions fit to training features. Fort et al. (2021) achieved state-of-the-art OOD detection performance on standard vision benchmarks by applying the MD to features extracted from large-scale, pre-trained vision transformers. González et al. (2021) applied the MD to medical imaging segmentation architectures by fitting a Gaussian distribution to the bottleneck features of an nnU-Net architecture (Isensee et al., 2020). Sun et al. (2022) replaced MD with a k-th nearest neighbor distance to relax the Gaussian assumption on the feature space. Ghosal et al. (2024) proposed Subspace Nearest Neighbor (SNN), a k-th nearest neighbor approach that reduces the dimensionality of the calculation by masking out irrelevant features. Karimi and Gholipour (2023) found that the Euclidean distance between the spectral features of a test image and its nearest neighbor in the training dataset achieved the best OOD detection performance on medical imaging segmentation tasks. Sastry and Oore (2020) measured the deviation of test images from training images by applying high-order Gram matrices to all neural network layers. Our work builds upon past research by improving the performance of MD through dimensionality reduction and comparing MD to KNN in a medical imaging feature space.

Hendrycks and Gimpel (2017) proposed Maximum Softmax Probability (MSP) as an OOD detection baseline, where OOD samples were identified by their prediction probabilities. The intuition behind this approach is that models should express more confidence on ID samples than on OOD samples. However, in practice, neural networks tend to express high confidence, even on OOD samples (Nguyen et al., 2015). Guo et al. (2017) calibrated model confidence, or aligned prediction probabilities with true correctness likelihoods, by scaling model logits with a single parameter (temperature scaling). Liang et al. (2018) utilized temperature scaling and small input perturbations to improve the performance of MSP (ODIN). Liu et al. (2020) furthered performance by replacing the softmax function with an energy function. Sun et al. (2021) outperformed previous methods by truncating hidden activations (ReAct). In this work, we compare MD and KNN to MSP, temperature scaling, and energy scoring.

OOD detection can also be performed using prediction uncertainties. While Bayesian neural networks provide probability distributions over weights, they are intractable for uncertainty estimation as they require significant modifications to neural network architectures and are computationally prohibitive. MC Dropout, which computes prediction uncertainties by combining multiple stochastic passes through a network at inference, was proposed by Gal and Ghahramani (2016) as a Bayesian approximation to Gaussian processes. Gal et al. (2017) further introduced concrete dropout, which improved uncertainty calibration by allowing for automated tuning of the dropout probability. Teye et al. (2018) approximated Bayesian inference with batch normalization. Lakshminarayanan et al. (2017) took a frequentist approach by ensembling neural networks to improve predictive uncertainty. While ensembling is computationally expensive, it performs well in medical imaging segmentation literature (Jungo et al., 2020; Mehrtash et al., 2020; Adams and Elhabian, 2023). To reduce the computational complexity of ensembling, Wen et al. (2020) proposed BatchEnsemble, which enables weight sharing. In this work, we compare MD and KNN to MC Dropout and ensembling.

3. Methods

3.1 Segmentation

We used six liver segmentation models (Table 1) that were based on six datasets (Table 2) for OOD detection analysis. The first model was the Swin UNETR (Tang et al., 2022) from Woodland et al. (2023) (hereafter called MRI UNETR). This model was trained on 337 T1-weighted liver MRI exams from the Duke Liver (DLDS) (Macdonald et al., 2020, 2023), Abdominal Multi-Organ Segmentation (AMOS) (Ji, 2022; Ji et al., 2022), and Combined Healthy Abdominal Organ Segmentation (CHAOS) (Kavur et al., 2019, 2020, 2021) datasets (collectively called MRI_{Tr}) and tested on 27 T1-weighted liver MRIs from The University of Texas MD Anderson Cancer Center (called MRI_{Te}). All MD Anderson images were retrospectively acquired under an internal review board (IRB)-approved protocol.

Table 1: Description of the segmentation models.

Name	Train	Test	Description
MRI UNETR	$\mathrm{MRI}_{\mathrm{Tr}}$	$\mathrm{MRI}_{\mathrm{Te}}$	UNETR from Woodland et al. (2023)
MRI Dropout	$\mathrm{MRI}_{\mathrm{Tr}}$	$\mathrm{MRI}_{\mathrm{Te}}$	UNETR with dropout enabled and predictions combined
MRI Ensemble	$\mathrm{MRI}_{\mathrm{Tr}}$	$\mathrm{MRI}_{\mathrm{Te}}$	Ensemble of 5 UNETRs
MRI+ UNETR	$\mathrm{MRI}+_{\mathrm{Tr}}$	$\mathrm{MRI}+_{\mathrm{Te}}$	UNETR from Patel et al. (2024)
MRI+ nnU-net	$\mathrm{MRI}+_{\mathrm{Tr}}$	$\mathrm{MRI}+_{\mathrm{Te}}$	nnU-net from Patel et al. (2024)
CT nnU-net	$\mathrm{CT}_{\mathrm{Tr}}$	$\mathrm{CT}_{\mathrm{Te}}$	nnU-net from MD Anderson

The next two models were created to enable comparison of MC Dropout and ensembling to MD and KNN. The second model, called MRI Dropout, was a Swin UNETR trained on $\mathrm{MRI}_{\mathrm{Tr}}$ with a 20% dropout rate and tested on $\mathrm{MRI}_{\mathrm{Te}}$. Enabling dropout added dropout layers into many architecture components, including after the positional embedding, the window-based multi-head self-attention module, and the multi-layer perceptron. The segmentation map for this model was generated by averaging the predictions obtained from enabling dropout during inference and conducting five forward passes per image. The third model, called MRI Ensemble, was an ensemble of five Swin UNETRs trained on $\mathrm{MRI}_{\mathrm{Tr}}$. The predictions from each of these five models were averaged to generate a final segmentation map.

Table 2: Description of the datasets.

Name	Description
$\mathrm{MRI}_{\mathrm{Tr}}$	337 T1-weighted abdominal MRIs from DLDS, AMOS, and CHAOS datasets
$\mathrm{MRI}_{\mathrm{Te}}$	27 T1-weighted abdominal MRIs from MD Anderson
$\mathrm{MRI}+_{\mathrm{Tr}}$	371 T1-weighted abdominal MRIs from MD Anderson and curated DLDS, AMOS, CHAOS, ATLAS datasets
$\mathrm{MRI}+_{\mathrm{Te}}$	352 T1-weighted abdominal MRIs from Houston Methodist
$\mathrm{CT}_{\mathrm{Tr}}$	2,840 abdominal CTs from MD Anderson
$\mathrm{CT}_{\mathrm{Te}}$	248 abdominal CTs from MD Anderson and BTCV challenge

The encoders of MRI UNETR, MRI Dropout, and MRI Ensemble models were pre-trained using self-distilled masked imaging (SMIT) (Jiang et al., 2022) with 3,610 unlabeled head and neck computed tomography scans (CTs) from the Beyond the Cranial Vault (BTCV) Segmentation Challenge dataset (Landman et al., 2015). The official Swin UNETR codebase, built on top of the Medical Open Network for AI (MONAI) (Consortium, 2021), was utilized for the pre-trained weights and training. Models were trained with default parameters for 1,000 epochs with the default batch size of 1. Each model was trained on a single node of a Kubernetes cluster containing eight A100 graphic processing units (GPUs) with 40 gigabytes (GB) of memory. A total of 100 GB of memory was requested from the cluster. Final model weights were selected according to the weights with the highest validation Dice similarity coefficient (DSC).

The rest of the liver segmentation models were previously trained at MD Anderson and were utilized for post-hoc OOD detection analysis. The fourth model, called MRI+

UNETR, builds upon MRI UNETR by expanding and curating the training and testing datasets. MRI+ UNETR was trained on 48 scans from the AMOS dataset, 172 scans from the DLDS dataset, 38 scans from the CHAOS dataset, 44 scans from the Tumor and Liver Automatic Segmentation (ATLAS) dataset (Quinton et al., 2023), and 69 scans from MD Anderson, for a total of 371 T1-weighted liver MRIs (collectively named MRI+$_{\mathrm{Tr}}$). 352 scans from 71 patients with hepatocellular carcinoma collected from Houston Methodist Hospital (called MRI+$_{\mathrm{Te}}$) were used for evaluation. Inclusion criteria for MRI+$_{\mathrm{Tr}}$ and MRI+$_{\mathrm{Te}}$ necessitated the entire liver to be visible, no prior liver surgery, and sufficient image quality to ensure the boundary of the liver was identifiable without pre-existing contours. The fifth model, named MRI+ nnU-net, was an nnU-net trained and tested on MRI$_{\mathrm{Tr}}$ and MRI+$_{\mathrm{Te}}$ that was included to enable a comparison between the nnU-net and Swin UNETR architectures. For more information on the MRI+ models, please refer to Patel et al. (2024).

The final model, named CT nnU-net, was an nnU-net trained on 2,840 internally obtained abdominal computed tomography (CT) scans (CT$_{\mathrm{Tr}}$) and tested on 248 CT scans (CT$_{\mathrm{Te}}$). It was included to expand our analysis to computed tomography. The training scans varied in the presence of and phase of contrast (portal-venous and arterial phases), states of liver disease and histology, presence of artifacts (including ablation needles, stents, and post-resection clips), and therapy stage (planning, intra-operative, and post-operative). 30 of the test scans came from the BTCV challenge (Landman et al., 2015), while the rest were acquired internally from MD Anderson. Images from an ongoing liver ablation clinical trial[1] and Anderson et al. (2021) were used in both the training and testing phases of the segmentation model.

Segmentation performance was evaluated with DSC, maximum Hausdorff distance (HD), and Normalized Surface Dice (NSD) with a threshold of 2 millimeters. One-sided paired t-tests were conducted with a significance level of α=.05 to determine the significance of performance improvements.

3.2 OOD Detection

To evaluate the detection of images that a segmentation model will perform poorly on, each model's test data was split into in-distribution (ID) and OOD categories based on that specific model's performance. An image was labeled ID for a model if the image had an associated DSC of at least 95%. Accordingly, an image was labeled OOD if it had a DSC under 95%. If there were not at least two ID images, the threshold was lowered to 80%. While we consider 80% DSC to be acceptable for clinical deployment, we prefer a 95% DSC as these contours are unlikely to require any editing in the clinical process. In practice, the threshold should be determined by the individual use case. For robustness, experiments were computed for 95%, 80%, and median value thresholds. Furthermore, DSCs were plotted against OOD scores to visually demonstrate how the results would change if the threshold changed.

Performance was measured with the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), the false positive rate at a 90% true positive rate (FPR90), and the amount of time in seconds it took to compute

[1] `https://www.clinicaltrials.gov/study/NCT04083378`

the OOD scores, with OOD considered the positive class. Averages and standard deviations (SDs) were reported across five OOD score calculations, with each score being calculated on the entire test dataset with a different NumPy random seed. t-tests were performed with a significance level $\alpha = 0.10$ to determine the significance of configuration improvements. All calculations were performed on a node of a Kubernetes cluster with 16 central processing units (CPUs) and a requested 256 megabytes of memory, with a 256 GB limit.

3.2.1 Mahalanobis distance

The Mahalanobis distance D measures the distance between a point x and a distribution with mean μ and covariance matrix Σ, $D^2 = (x - \mu)^T \Sigma^{-1} (x - \mu)$ (Mahalanobis, 1936). Lee et al. (2018) first proposed using the MD for OOD detection to calculate the distance between test images embedded by a classifier and a Gaussian distribution fit to class-conditional embeddings of the training images. Similarly, González et al. (2021) used the MD for OOD detection in segmentation networks by extracting embeddings from the encoder of an nnU-Net. As distances in high dimensions are subject to the curse of dimensionality, both sets of authors decreased the dimensionality of the embeddings through average pooling. Lee et al. (2018) suggested pooling the features such that the height and width dimensions of the features become singular. González et al. (2021) pooled features with a kernel size and stride of 2 until the feature dimensionality fell below 10,000.

The MD was calculated on features extracted from the bottlenecks of the liver segmentation models. For the UNETRs, the projected features `x` were saved from the `Trans_Unetr` class. For the nnU-nets, the features `skips` were saved from the `PlainConvUNet` class for each sliding window and subsequently concatenated. As the nnU-nets automatically cropped their inputs, the size of the bottleneck feature dimension representing the number of concatenated sliding windows could not be standardized. To account for this, average pooling was applied across the concatenated embeddings such that the size of the dimension representing the number of concatenated windows was made singular. The size of the bottleneck features for the UNETR models was standardized by resizing model inputs to (256, 128, 128) prior to feature extraction. After extraction, all features were flattened to prepare them for distance calculations. Note that nnU-net uses instance normalization in the encoding process, whereas UNETR uses layer normalization.

Gaussian distributions were fitted on the raw training embeddings. For each test embedding, the MD between the embedding and the corresponding Gaussian distribution was calculated and used as the OOD score. Covariance matrices were estimated empirically with maximum likelihood estimation.

3.2.2 K-Nearest Neighbors

When Lee et al. (2018) introduced MD for OOD detection, the authors showed that the posterior distribution of a softmax classifier can be modeled by a generative classifier defined by Gaussian Discriminant Analysis, thereby demonstrating that classifier embeddings can be Gaussian-distributed. However, this does not guarantee the embeddings are Gaussian-distributed, as demonstrated by classifier embeddings failing normality tests (Sun et al., 2022). Additionally, this analysis has not been extended to segmentation networks.

Sun et al. (2022) first proposed using a k-th nearest neighbor distance as a non-parametric alternative to the MD. In their work, KNN improved overall performance over MD across five OOD detection benchmark datasets, though performance improvements were dataset-dependent. In this work, we propose KNN for medical imaging segmentation networks. We define the k-th nearest neighbor distance to be the Euclidean distance between a test embedding $f(x)$ and its k-th nearest training embedding $f(z_k)$ where f is a trained U-Net encoder

$$||f(x) - f(z_k)||_2.$$

This k-th nearest neighbor distance serves as the OOD score. A hyperparameter search was performed over k such that $k \in \{2, 4, 8, 16, 32, 64, 128, 256\}$. KNN was performed on the features that were extracted for MD.

3.2.3 Dimensionality Reduction

As distances in extremely high-dimensional spaces often lose meaning (Aggarwal et al., 2001), experiments were performed on the effect of decreasing the size of the bottleneck features using average pooling, PCA, UMAP (McInnes et al., 2020), and t-distributed stochastic neighbor embeddings (t-SNE) (Van der Maaten and Hinton, 2008). For average pooling, features were pooled in both 2- and 3-dimensions with kernel size k and stride s for $(k, s) \in \{(2, 1), (2, 2), (3, 1), (3, 2), (4, 1)\}$. For PCA, each embedding was flattened and standardized. For both PCA and UMAP, a hyperparameter search was performed over the number of components n such that $n \in \{2, 4, 8, 16, 32, 64, 128, 256\}$. The PyTorch, scikit-learn, and UMAP Python packages were used for the dimensionality reduction (McInnes et al., 2020). Outside of the hyperparameter searches mentioned above, default parameters were used. A visual representation of integrating dimensionality reduction with MD and KNN is available in Figure 1.

The efficacy of PCA, t-SNE, and UMAP for OOD detection was examined qualitatively by plotting the features from the MRI UNETR reduced to two dimensions. Features generated by the MRI_{Tr} and MRI_{Te} datasets (split into ID and OOD by 95% DSC) were plotted and subsequently compared. Additionally, MRI_{Te} embeddings were plotted by DSC, and MRI_{Tr} embeddings were plotted by source (DLDS, AMOS, and CHAOS).

3.2.4 Comparison Methods

We used MSP (Hendrycks and Gimpel, 2017) as a post-hoc detection baseline. In addition, we evaluated two proposed improvements to MSP: temperature scaling (Guo et al., 2017) and energy scoring (Liu et al., 2020). To provide non-post-hoc detection baselines, MD and KNN were further compared with MC Dropout (Gal and Ghahramani, 2016) and ensembling (Lakshminarayanan et al., 2017) on the MRI_{Te} dataset.

To undertake the OOD detection task using MSP, the logits pertaining to the foreground and background classes were stacked, followed by a softmax. The maximum probability for each voxel was then calculated across the foreground and background classes. The average of these maximum probabilities for the entire image was then computed. We subtracted this average from 1 to get the final OOD score for MSP. For temperature scaling, logits were divided evenly by $T \in \{2, 3, 4, 5, 10, 100, 1000\}$ before the softmax was applied. For

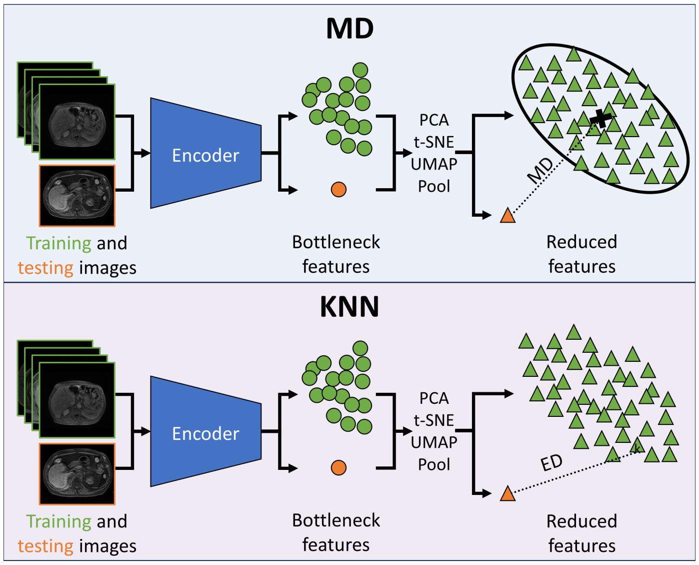


Figure 1: MD and KNN pipelines with dimensionality-reduced features using either PCA, t-SNE, UMAP, or average pooling (Pool). The encoder is a trained encoder from a U-Net architecture. k is the k-th nearest neighbor.

energy scoring, the OOD score was calculated as

$$E(x; f) = -T * \log \sum_{i=1}^{k} \frac{f_i(x)}{T}$$

for image x, i-th logit f_i, and scaling parameter $T \in \{1, 2, 3, 4, 5, 10, 100, 1000\}$. For MC Dropout, voxel-wise standard deviations were calculated across the predictions from the five forward passes of MRI Dropout with dropout enabled. The average of these standard deviations was used as the MC Dropout OOD score. Similarly, the average of voxel-wise standard deviations calculated across the predictions from the five members of MRI Ensemble was used as the ensembling OOD score.

To further evaluate the performance of the OOD detection methods on the poor segmentation performance task, Pearson correlation coefficients (PCCs) were computed between the OOD scores of the best-performing configuration of each OOD detection method and the DSC, HD, and NSD segmentation metrics with a significance level of $\alpha = .10$. These relationships were further explored qualitatively by plotting the OOD scores against DSCs.

4. Results

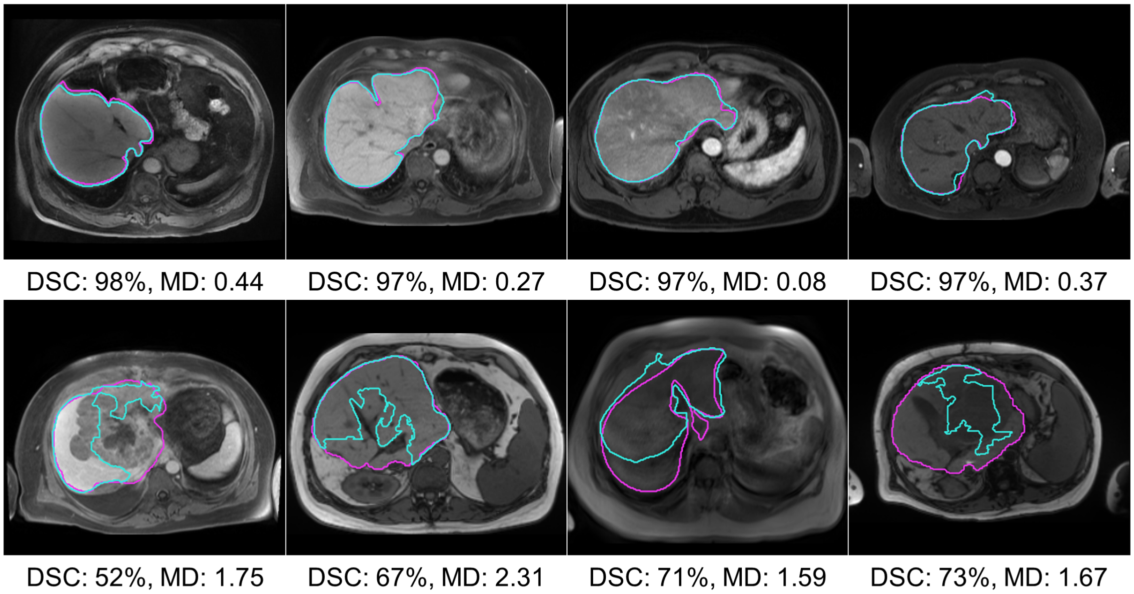
4.1 Segmentation

The segmentation performance of the MRI Dropout and MRI Ensemble models improved on that of the MRI UNETR by averaging the predictions (paired t-tests for HD, $p = .013$ Dropout, $p = .020$ Ensemble; Table 3). Thresholds for OOD detection were set at 95%

DSC for all models except the MRI+ UNETR. This model only produced one DSC over 95%, so the threshold was lowered to 80%. 13 images determined to be OOD were shared across the MRI UNETR, MRI Dropout, and MRI Ensemble models. MRI+ UNETR and MRI+ nnU-net performed similarly on MRI+_{Te}, with MRI+ UNETR achieving a lower HD and MRI+ nnU-net achieving a higher NSD (paired t-tests, $p < .001$ all tests). Figure 2 displays visual examples of the segmentation quality of the MRI UNETR.

Table 3: Average (±SD) segmentation performances. # OOD refers to the number of test images determined to be OOD. Arrows denote whether higher or lower is better. Bold highlights the best performance per test dataset, with underlined performances denoting statistical significance.

Model	**DSC** (±SD) ↑	**HD** (±SD) ↓	**NSD** (±SD) ↑	# OOD
MRI UNETR	0.89 (±0.11)	34.38 (±25.67)	0.72 (±0.22)	14
MRI Dropout	**0.91** (±0.09)	**23.30** (±24.63)	0.75 (±0.20)	14
MRI Ensemble	**0.91** (±0.10)	24.23 (±24.94)	**0.76** (±0.20)	13
MRI+ UNETR	**0.90** (±0.04)	**18.10** (±10.66)	0.65 (±0.10)	7
MRI+ nnU-net	**0.90** (±0.03)	31.27 (±25.06)	**0.78** (±0.08)	349
CT nnU-net	**0.97** (±0.01)	**23.10** (±26.34)	**0.96** (±0.05)	22



DSC: 98%, MD: 0.44

DSC: 97%, MD: 0.27

DSC: 97%, MD: 0.08

DSC: 97%, MD: 0.37

DSC: 52%, MD: 1.75

DSC: 67%, MD: 2.31

DSC: 71%, MD: 1.59

DSC: 73%, MD: 1.67

Figure 2: Segmentations with high (top) and low (bottom) DSCs along with their corresponding MDs, calculated in conjunction with PCA with two components. Pink is the ground truth segmentation; teal is the MRI UNETR segmentation.

4.2 OOD Detection

4.2.1 Mahalanobis Distance

The calculation of the MD on raw bottleneck features was computationally intractable (Table 4). For the MRI UNETR, the inverse of the covariance matrix took ∼2.6 hours to compute. Once computed, it took 75.5 GB to store the inverse. Once the matrix was in memory, each MD calculation took ∼2 seconds. The average (± SD) MD on MRI_{Tr} was 1203.02 (± 24.66); whereas, the average (± SD) MD on MRI_{Te} was $1.47{\times}10^9$ $(\pm 8.66 \times 10^8)$ and $1.52 \times 10^9 (\pm 9.10 \times 10^8)$ for ID and OOD images, respectively.

Table 4: MD-based OOD detection results. Results are based on a 95% DSC threshold for all models except the MRI+ UNETR, where the threshold was set at 80% DSC. Only the best-performing configuration by AUROC is reported for each dimensionality reduction technique (Reduct): no reduction (None), PCA(n_p) with n_p components, t-SNE, UMAP(n_u) with n_u components, and n_d-dimensional average pooling with kernel size k and stride s, Pooln_dD(k,s). Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether higher or lower is better. Bold highlights the best performance per model, with underlined performances denoting statistical significance. Appendix A.1 contains the results with varied thresholds, and Appendix B.1 contains the full hyperparameter searches.

Model	**Reduct**	**AUROC ↑**	**AUPRC ↑**	**FPR90 ↓**	**Seconds ↓**
MRI UNETR	None	0.48 (±0.00)	0.61 (±0.00)	1.00 (±0.00)	9,354.34 (±48.53)
	PCA(256)	$\underline{\textbf{0.93}}$ (±0.00)	$\underline{\textbf{0.94}}$ (±0.00)	$\underline{\textbf{0.23}}$ (±0.00)	$\underline{\textbf{2.82}}$ (±0.14)
	t-SNE	0.70 (±0.08)	0.72 (±0.12)	0.71 (±0.14)	4.70 (±0.28)
	UMAP(2)	0.77 (±0.08)	0.79 (±0.11)	0.57 (±0.08)	10.44 (±0.22)
	Pool2D(3,2)	0.82 (±0.00)	0.86 (±0.00)	0.46 (±0.00)	15.32 (±8.92)
MRI+ UNETR	None	0.53 (±0.00)	0.03 (±0.00)	0.78 (±0.00)	10,070.78 (±141.69)
	PCA(32)	$\underline{\textbf{0.85}}$ (±0.01)	$\underline{\textbf{0.13}}$ (±0.00)	$\underline{\textbf{0.34}}$ (±0.02)	$\underline{\textbf{1.86}}$ (±0.32)
	t-SNE	0.66 (±0.00)	0.03 (±0.00)	0.45 (±0.00)	5.77 (±0.06)
	UMAP(2)	0.68 (±0.07)	0.05 (±0.01)	0.49 (±0.06)	21.29 (±0.43)
	Pool2D(4,1)	0.64 (±0.00)	0.04 (±0.00)	0.64 (±0.00)	16.08 (±13.41)
MRI+ nnU-net	None	0.69 (±0.00)	$\underline{\textbf{1.00}}$ (±0.00)	0.67 (±0.00)	4,125.96 (±13.12)
	PCA(8)	$\underline{\textbf{0.96}}$ (±0.00)	$\underline{\textbf{1.00}}$ (±0.00)	$\underline{\textbf{0.00}}$ (±0.00)	$\underline{\textbf{1.12}}$ (±0.04)
	t-SNE	0.70 (±0.18)	$\underline{\textbf{1.00}}$ (±0.00)	0.87 (±0.16)	4.72 (±0.11)
	UMAP(16)	0.82 (±0.08)	$\underline{\textbf{1.00}}$ (±0.00)	0.67 (±0.03)	19.07 (±0.96)
	Pool2D(2,1)	0.85 (±0.00)	$\underline{\textbf{1.00}}$ (±0.00)	0.67 (±0.00)	1,579.73 (±52.41)
CT nnU-net	None	0.41 (±0.00)	0.10 (±0.00)	0.91 (±0.00)	5,856.21 (±63.04)
	PCA(32)	0.56 (±0.00)	0.17 (±0.00)	0.92 (±0.00)	$\underline{\textbf{8.17}}$ (±0.23)
	t-SNE	0.59 (±0.04)	0.20 (±0.02)	0.80 (±0.02)	13.75 (±0.35)
	UMAP(128)	$\underline{\textbf{0.68}}$ (±0.03)	**0.22** (±0.01)	**0.74** (±0.09)	288.42 (±25.99)
	Pool2D(2,2)	0.59 (±0.00)	0.13 (±0.00)	0.84 (±0.00)	163.84 (±19.54)

4.2.2 K-Nearest Neighbors

While the MD calculation was not tractable on raw features, the KNN calculation was (Table 5). The calculation took ∼0.02 seconds per image for the MRI data and ∼0.08 seconds per image for the liver CT data. In addition to being more scalable, KNN improved the AUROC over the MD applied to raw features for all models (t-tests, $p < .001$ all tests).

Table 5: KNN-based OOD detection of poor performance results. Results are based on a 95% DSC threshold for all models except the MRI+ UNETR, where the threshold was set at 80% DSC. Only the best-performing configuration by AUROC is reported for each dimensionality reduction technique (Reduct): no reduction (None), PCA(n_p) with n_p components, t-SNE, UMAP(n_u) with n_u components, and n_d-dimensional average pooling with kernel size k and stride s, Pooln_dD(k,s). Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether higher or lower is better. Bold highlights the best performance, with underlined performances denoting statistical significance. Appendix A.2 contains the results with varied thresholds, and Appendix B.2 contains the link to the full hyperparameter searches.

Model	**Reduct**	**K**	**AUROC ↑**	**AUPRC ↑**	**FPR90 ↓**	**Seconds ↓**
MRI UNETR	None	256	0.87 (±0.00)	0.88 (±0.00)	0.31 (±0.00)	**0.78** (±0.00)
	PCA(2)	256	0.90 (±0.00)	0.92 (±0.00)	0.31 (±0.00)	0.95 (±0.05)
	t-SNE	256	0.77 (±0.05)	0.83 (±0.04)	0.74 (±0.06)	4.48 (±0.07)
	UMAP(32)	256	0.83 (±0.05)	0.85 (±0.04)	0.51 (±0.08)	6.70 (±0.15)
	Pool2D(3,1)	256	**0.94** (±0.00)	**0.95** (±0.00)	**0.23** (±0.00)	0.90 (±0.00)
MRI+ UNETR	None	256	0.76 (±0.00)	0.25 (±0.00)	0.59 (±0.00)	9.44 (±0.16)
	PCA(32)	64	0.84 (±0.01)	0.17 (±0.02)	0.40 (±0.01)	**1.57** (±0.06)
	t-SNE	64	0.70 (±0.00)	0.04 (±0.00)	**0.37** (±0.00)	6.49 (±0.19)
	UMAP(4)	64	0.78 (±0.05)	0.09 (±0.06)	0.42 (±0.07)	15.70 (±0.45)
	Pool3D(2,2)	256	**0.87** (±0.00)	**0.33** (±0.00)	0.43 (±0.00)	11.21 (±7.66)
MRI+ nnU-net	None	256	0.96 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	6.78 (±0.10)
	PCA(8)	256	0.97 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	1.16 (±0.07)
	t-SNE	128	0.67 (±0.08)	**1.00** (±0.00)	1.00 (±0.00)	4.76 (±0.11)
	UMAP(2)	2	0.96 (±0.04)	**1.00** (±0.00)	0.13 (±0.16)	14.91 (±2.08)
	Pool2D(2,2)	256	**0.98** (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	**0.74** (±0.04)
CT nnU-net	None	8	0.52 (±0.00)	0.13 (±0.00)	0.94 (±0.00)	37.64 (±0.67)
	PCA(8)	4	0.55 (±0.00)	0.15 (±0.00)	0.97 (±0.00)	**4.94** (±0.04)
	t-SNE	256	0.46 (±0.00)	0.19 (±0.00)	0.95 (±0.01)	12.28 (±0.09)
	UMAP(4)	256	**0.65** (±0.01)	**0.24** (±0.05)	**0.88** (±0.02)	199.93 (±1.96)
	Pool3D(2,2)	4	0.54 (±0.00)	0.14 (±0.00)	0.96 (±0.00)	81.96 (±26.79)

4.2.3 Dimensionality Reduction

Paired with MD, all dimensionality reduction techniques resulted in improvements in the AUROC (t-tests, $p = .003$ UMAP/MRI+ UNETR, $p < .001$ all other tests; Table 4). On the MRI models, PCA achieved the best performance, outperforming average pooling by 0.14 (±0.06)% AUROC and 535.11 (±903.70) seconds. For CT nnU-net, UMAP achieved the best AUROC, outperforming average pooling by 0.09. Figure 2 displays MDs computed on PCA-reduced features, along with the corresponding segmentations. In this figure, higher MDs were associated with poor segmentation performance.

Paired with KNN, PCA, and average pooling resulted in AUROC improvements for all models (t-tests, $p < .001$ all tests; Table 5). Similar to MD, KNN on UMAP-reduced features achieved the highest AUROC for the CT nnU-net (t-tests, $p < .001$ all tests). In contrast to MD, KNN applied to average pooled features achieved the highest AUROCs for the MRI models (t-tests, $p = .001$ UMAP/MRI UNETR, $p = .007$ UMAP/MRI+ UNETR, $p < .001$ all other tests except UMAP/MRI+ nnU-net). On the MRI models, KNN outperformed MD when applied to average pooled features by 0.16 (±0.05) AUROC and 532.76 (±739.81) seconds. Overall, KNN applied to average-pooled features slightly outperformed MD on PCA-reduced features by 0.02 (±0.00) AUROC for the MRI models (t-tests, $p = .003$ MRI+ UNETR, $p < .001$ all other tests).

Figure 3 visualizes the 2D embeddings produced by PCA, t-SNE, and UMAP for the MRI UNETR. In addition, covariance ellipses generated by the training distribution are plotted, representing one and two standard deviations away from the mean training embedding. PCA mapped most ID test images within one standard deviation of the mean training embedding (the image not within the first deviation contained a motion artifact). On the other hand, most OOD test images were mapped outside of the first standard deviation. When test embeddings were visualized by their DSC, the three reduction techniques mapped the images with the lowest DSCs farthest from the mean training embedding. Moreover, all three techniques clustered training embeddings by source. The 26 images from the AMOS dataset mapped outside the second standard deviation by all techniques were deemed to be of low perceptual resolution by a physician. These were the only images from the AMOS dataset whose axial dimension was larger than the sagittal dimension. Sample images from both AMOS clusters are shown in Figure 5 in Appendix C.

4.2.4 Comparison Methods

On MRI_{Te}, MSP, MC Dropout, and ensembling outperformed MD and KNN (t-tests on AUROC, $p < .001$ all tests), with MC Dropout perfectly differentiating between ID and OOD categories (Table 6). Similarly, MSP outperformed MD and KNN on CT_{Te} (t-tests on AUROC, $p = .036$ MD, $p < .001$ KNN). In contrast, MD and KNN outperformed the output-based methods on MRI+_{Te} (t-tests on AUROC, $p < .001$ all tests).

Although originally intended to improve the performance of MSP, both temperature scaling and energy scoring performed worse than MSP on the MRI_{Te} and CT_{Te} datasets (t-tests on AUROC, $p < .001$ all tests). Overall, MSP achieved the highest AUROCs for the MRI UNETR and CT nnU-net models (t-tests, $p = .036$ MD/CT nnU-net, $p < .001$ all other tests), and KNN achieved the highest AUROCs for the MRI+ UNETR and nnU-net models (t-tests, $p = .003$ MD/MRI+ UNETR, $p < .001$ all other tests).

The OOD scores from KNN, MSP, and temperature scaling were significantly correlated with DSC across all models (Table 7). MD was significantly correlated with DSC for all the MRI models. Energy scoring was significantly correlated with DSC for only the nnU-nets.

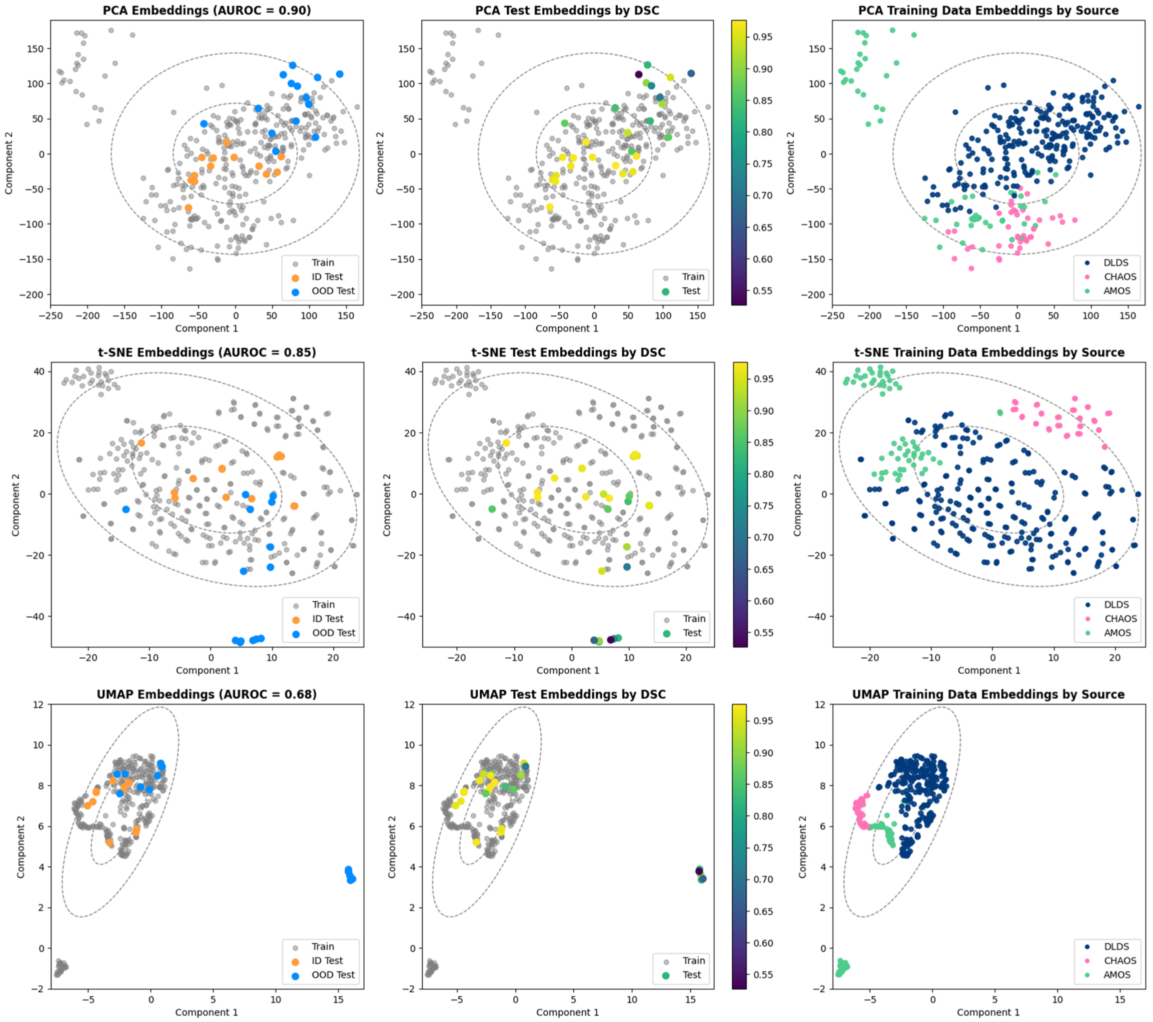


Figure 3: Visualization of 2D projections of MRI UNETR embeddings. (Top Row) PCA projections. (Middle Row) t-SNE projections. (Bottom Row) UMAP projections. (Left Column) Test projections split into ID and OOD by 95% DSC. (Middle Column) Test projections by DSC. (Right Column) Projections for the training data by source. The gray ellipses are the covariance ellipses (one and two standard deviations) for the training distribution.

MC Dropout and ensembling were significantly correlated with all segmentation metrics on MRI_{Te}, notably achieving PCCs of 0.96 and 0.97 with HD ($p < .001$ all correlations). Considering only post-hoc detection methods, MSP achieved the best correlations with DSC for the MRI UNETR and CT nnU-net models (-0.77 and -0.28) and the best correlations with HD and NSD for MRI+ nnU-net (0.30 and -0.31; t-tests, $p < .001$ all tests). MD and temperature scaling achieved the best correlations with DSC for the MRI+ UNETR and MRI+ nnU-net models, respectively (-0.14 and -0.31; t-tests, $p < .001$ all tests).

Table 6: OOD detection results for the best-performing configurations of the comparison methods: MSP, temperature scaling (TS) and energy scoring (Energy) with temperature T, MC Dropout (MCD), and ensembling (Ensemble). The results are averages (±SD) across 5 runs. Arrows denote whether a higher or lower value is better. Bold highlights the best performance per model, with underlined performances denoting statistical significance. Appendix A.3 contains the results with varied thresholds, and Appendix B.3 contains the full hyperparameter searches.

Model	**Method**	**T**	**AUROC ↑**	**AUPRC ↑**	**FPR90 ↓**	**Seconds ↓**
MRI UNETR	MSP	-	**0.96** (±0.00)	**0.97** (±0.00)	**0.00** (±0.00)	8.78 (±0.11)
	TS	2	0.90 (±0.00)	0.93 (±0.00)	0.15 (±0.00)	9.25 (±0.11)
	Energy	3	0.55 (±0.00)	0.57 (±0.00)	0.69 (±0.00)	**7.08** (±0.14)
MRI Dropout	MCD	-	**1.00** (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	**14.56** (±0.20)
MRI Ensemble	Ensemble	-	**0.96** (±0.00)	**0.96** (±0.00)	**0.14** (±0.00)	**14.02** (±0.05)
MRI+ UNETR	MSP	-	0.57 (±0.00)	**0.09** (±0.00)	**0.59** (±0.00)	58.70 (±0.48)
	TS	2	0.47 (±0.00)	0.05 (±0.00)	0.78 (±0.00)	60.57 (±0.31)
	Energy	1000	**0.61** (±0.00)	0.03 (±0.00)	0.71 (±0.00)	**35.86** (±0.41)
MRI+ nnU-net	MSP	-	0.45 (±0.00)	0.99 (±0.00)	**1.00** (±0.00)	1,114.34 (±0.51)
	TS	10	0.55 (±0.00)	0.99 (±0.00)	**1.00** (±0.00)	1,252.78 (±0.73)
	Energy	10	**0.61** (±0.00)	**1.00** (±0.00)	**1.00** (±0.00)	**186.88** (±0.47)
CT nnU-net	MSP	-	**0.72** (±0.00)	**0.29** (±0.00)	**0.57** (±0.00)	568.51 (±0.37)
	TS	3	0.68 (±0.00)	0.26 (±0.00)	0.69 (±0.00)	699.07 (±0.37)
	Energy	2	0.67 (±0.00)	0.26 (±0.00)	0.75 (±0.00)	**105.66** (±0.60)

Figure 4 plots OOD scores against DSCs. By moving the horizontal line vertically, one can visualize how the OOD detection performance would change if the DSC threshold were changed. MSP, ensembling, and MC Dropout visually demonstrated the strongest negative linear relationship between OOD scores and DSC. KNN and MD assigned noticeably higher OOD scores to six images with a wide range of DSCs. These images came from the same patient who had a large tumor in the liver, resulting in missing liver segments (Figure 6, Appendix D). The training-based methods assigned the noticeably highest OOD score to a scan with an imaging artifact. Instead of providing the intended further separation of softmax score distributions, temperature scaling and energy scoring visually pushed the distributions closer together.

5. Discussion

Our work provides several key takeaways. First, MD is highly sensitive to the methodology used to reduce the feature space. Past research reduced feature dimensionality with average pooling with fixed parameters (Lee et al., 2018; González et al., 2021). Our work demonstrates that this practice may not achieve the best results, considering average pooling was outperformed by PCA for the MRI data and UMAP for the CT data. While PCA and

Figure 4: OOD scores plotted against DSC for $\mathrm{MRI}_{\mathrm{Te}}$. Horizontal lines represent 95% DSC. Vertical lines represent the 90% TPR.

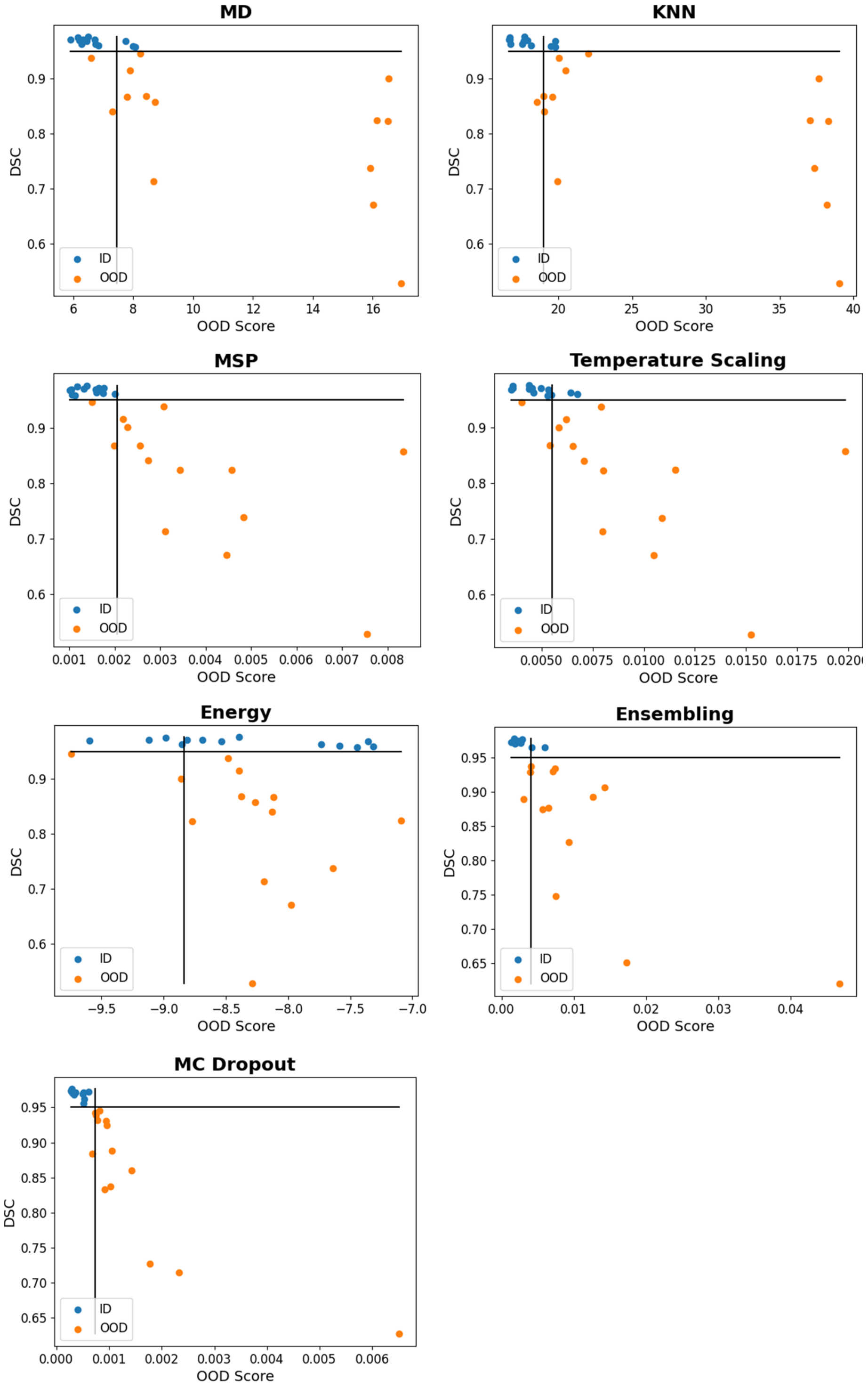


Table 7: Correlation results for the best-performing configuration of all methods: MD and KNN with dimensionality-reduction techniques PCA(n) and UMAP(n) with n components and n_d-dimensional average pooling with kernel size k and stride s Pooln_dD(k,s), MSP, temperature scaling (TS) and energy scoring (Energy) with temperature T, MC Dropout (MCD), and ensembling (Ensemble). * represents that each of the five correlation coefficients that were averaged over were statistically significant. Arrows denote whether a higher or lower value is better. Bold highlights the best performance per model, with underlined performances denoting statistical significance.

Model	Method	Config	PCC [DSC] ↓	PCC [HD] ↑	PCC [NSD] ↓
MRI UNETR	MD	PCA(256)	-0.74* (±0.00)	0.09 (±0.00)	**-0.76*** (±0.00)
	KNN	Pool2D(3,1) K=256	-0.72* (±0.00)	0.05 (±0.00)	-0.73* (±0.00)
	MSP	-	**-0.77*** (±0.00)	0.53* (±0.00)	-0.70* (±0.00)
	TS	T=2	-0.69* (±0.00)	**0.58*** (±0.00)	-0.64* (±0.00)
	Energy	T=3	-0.22 (±0.00)	0.20 (±0.00)	-0.24 (±0.00)
MRI Dropout	MCD	-	**-0.86*** (±0.00)	**0.96*** (±0.00)	**-0.67*** (±0.00)
MRI Ensemble	Ensemble	-	**-0.82*** (±0.00)	**0.97*** (±0.00)	**-0.63*** (±0.00)
MRI+ UNETR	MD	PCA(32)	**-0.14*** (±0.00)	0.05 (±0.00)	0.02 (±0.00)
	KNN	Pool3D(2,2) K=256	-0.13* (±0.00)	0.03 (±0.00)	**-0.03** (±0.00)
	MSP	-	-0.13* (±0.00)	-0.01 (±0.00)	0.10 (±0.00)
	TS	T=2	-0.13* (±0.00)	-0.02 (±0.00)	0.12 (±0.00)
	Energy	T=1000	-0.04 (±0.00)	**0.09** (±0.00)	**-0.03** (±0.00)
MRI+ nnU-net	MD	PCA(8)	-0.20* (±0.00)	0.11* (±0.00)	0.12* (±0.00)
	KNN	Pool2D(2,2) K=256	-0.27* (±0.00)	0.22* (±0.00)	0.14* (±0.00)
	MSP	-	-0.13* (±0.00)	**0.30*** (±0.00)	**-0.31*** (±0.00)
	TS	T=10	**-0.31*** (±0.00)	0.20* (±0.00)	0.04 (±0.00)
	Energy	T=10	-0.16* (±0.00)	-0.02 (±0.00)	0.07 (±0.00)
CT nnU-net	MD	UMAP(128)	-0.11 (±0.03)	0.05 (±0.04)	-0.10* (±0.02)
	KNN	UMAP(4) K=256	-0.21* (±0.01)	**0.23*** (±0.01)	-0.16* (±0.01)
	MSP	-	**-0.28*** (±0.00)	**0.23*** (±0.00)	**-0.29*** (±0.00)
	TS	T=3	-0.22* (±0.00)	0.20* (±0.00)	-0.25* (±0.00)
	Energy	T=2	-0.20* (±0.00)	0.19* (±0.00)	-0.23* (±0.00)

UMAP demonstrated promise, the best dimensionality reduction technique and parameter configuration is likely dataset and architecture-dependent. For example, while PCA and UMAP with only a few components performed well for liver segmentation, they may discard information important to the segmentation of smaller anatomical structures such as tumors. Architecture components, such as automatic cropping on an image-by-image basis, may also affect the applicability of a technique. Therefore, using a validation dataset to choose a dimensionality reduction technique and configuration could improve downstream OOD detection. Our findings align with those of Ghosal et al. (2024), whose work demonstrates that a careful and methodological selection of a subspace of features can improve feature-based OOD detection. A fundamental difference between our approach and that of González et al. (2021) is that we define OOD cases based on model performance. This difference could explain the decline in MD performance observed with average pooling in our study.

Second, raw and reduced segmentation model features may not be Gaussian-distributed, challenging the suitability of MD to all OOD detection tasks. This is especially true for medical imaging models, as the limited number of training images can lead to less clearly defined training distributions. Moreover, the Gaussian assumption of MD fails to account for the potential multi-modality of medical imaging distributions arising from various factors such as differences across source datasets, acquisition parameters, contrast phases, disease states, artifacts, and stages of therapy. In our work, a non-parametric approach, KNN, outperformed MD on raw features across all segmentation models and thresholds. In addition, MD outperformed KNN on average pooled features from the MRI segmentation models across all thresholds. Furthermore, our visualization of 2D embeddings highlighted significant gaps in the training distributions, with some training images positioned far from the central modes of the distributions. Our findings corroborate those of Sun et al. (2022), who demonstrated that parametric approaches are not always suitable for OOD detection.

Finally, the best OOD detection method for detecting poor segmentation performance may be task-dependent. In González et al. (2021), MD outperformed MC Dropout and MSP when applied to lung lesion segmentation. In González et al. (2022), MD and MSP achieved perfect differentiation for hippocampus segmentation, outperforming MC Dropout. For prostrate segmentation, on the other hand, MD and MC Dropout achieved perfect differentiation, outperforming MSP. In our liver segmentation study, MC Dropout outperformed MD, with MC Dropout achieving perfect differentiation. This suggests that MC Dropout may be better suited for tasks involving large and easily identifiable anatomical structures. Furthermore, in our liver segmentation study, MSP marginally outperformed MD on the test datasets drawn from MD Anderson, whereas MD substantially outperformed MSP on the test dataset from Houston Methodist. Moreover, on the MD Anderson test datasets, temperature scaling and energy scoring performed worse than MSP, a finding shared with González et al. (2021, 2022). However, energy scoring achieved the best performance of the output-based methods on the test dataset from Houston Methodist.

This work has several limitations. First, privacy must be considered when utilizing KNN, as the embeddings from all training images must be stored. Second, this work solely focused on liver segmentation. While this focus is advantageous for liver cancer research, the presented results may not extend to other anatomical structures. Third, the OOD detection thresholds relied on DSC, whereas surface-based metrics may better estimate whether a contour is clinically acceptable (Baroudi et al., 2023). Fourth, this work defined OOD detection based on model performance. Although this definition is important to consider for patient safety, the presented results may not extend to other OOD definitions. In addition, this definition caused the proportion of OOD test images to vary across models, limiting a direct comparison of OOD detection performance across these models. Finally, due to the automatic cropping nature of the nnU-nets utilized, all nnU-net embeddings had to be average pooled across the dimension representing the number of patches. Therefore, the nnU-net results are not a true representation of the distances applied to raw, PCA-, t-SNE-, and UMAP-reduced features.

Our work has several potential applications. First, a warning that the model likely failed could be added to automated segmentations with OOD scores above a specified threshold in a clinical setting. This would protect against automation bias, which would, in turn, protect patients whose scans have uncommon attributes. Ensembling and MC Dropout may be well suited for a liver segmentation task if the computational resources are available due to their superior segmentation and OOD detection performance. Second, detecting poor segmentation performance in retrospective studies where a large corpus of data is to be segmented. As reviewing all autosegmentations would be infeasible, human evaluators would only need to review the autosegmentations associated with large OOD scores. In this application, computational costs may outweigh performance. Accordingly, KNN may be advantageous to utilize. Third, the dimensionality techniques could provide a visualization tool for segmentation model creators to analyze how their model views their data. For example, using PCA with two components highlighted the images of low perceptual quality in the AMOS dataset. Lastly, this work could be used to diversify institutional training datasets by determining which images have the most utility to label. The OOD scores of scans in unlabeled institutional databases would elucidate the most challenging cases and the cases that differ the most from the original training dataset.

This research provides several avenues for future work. One of the biggest barriers to developing post-hoc OOD detection pipelines for medical imaging segmentation models is the number of choices one must consider when building their framework. Considering only feature-based methods for a moment, one must determine if they are going to use the features directly (Lee et al., 2018), a spectral analysis of the features (Karimi and Gholipour, 2023), or pairwise feature correlations with Gram matrices (Sastry and Oore, 2020). Then there are questions of which features should be used González et al. (2022); Anthony and Kamnitsas (2023), and if multiple are used, how to best aggregate them (Lee et al., 2018). Once the features are chosen, one must determine how to properly reduce them to satisfy computational requirements and optimize performance (Woodland et al., 2023; Ghosal et al., 2024). At this point, one should consider if parametric or non-parametric distances are the most appropriate for the reduced features (Sun et al., 2022). These considerations open up a plethora of avenues for future work. Research regarding each of these factors is of benefit to the field, in addition to large-scale application studies that demonstrate superior configurations for specific scenarios. However, what the field is most lacking is the collaborative infrastructure to automate these decision processes for specific models and validation datasets. In reality, the best configurations are most likely task-dependent, and most hospital systems developing segmentation models do not have the resources to perform such exhaustive searches.

6. Conclusion

In this work, MD was applied to dimensionality-reduced bottleneck features of a Swin UNETR trained for liver segmentation on T1-weighted MRIs. The resulting pipeline was able to embed entire 3D medical images into several components. These components were not only sufficient to cluster datasets drawn from different institutions but also could detect scans that the model performed poorly on with high performance and minimal computational cost (less than one second on CPUs). We validated our methods on previously trained liver segmentation models and found that either PCA or UMAP improved performance over average pooling for all models. Furthermore, we applied KNN to all models post hoc and found that it drastically outperformed the MD on raw and average pooled features: on a nnU-net trained on liver MRIs, it increased the AUROC to 96% from 69% and decreased the amount of time required to compute OOD scores for 352 MRIs from an hour to 7 seconds.

Acknowledgments

Research reported in this publication was supported in part by the Tumor Measurement Initiative through the MD Anderson Strategic Initiative Development Program (STRIDE), the Helen Black Image Guided Fund, the Image Guided Cancer Therapy Research Program at The University of Texas MD Anderson Cancer Center, a generous gift from the Apache Corporation, and the National Cancer Institute of the National Institutes of Health under award numbers R01CA221971, R01CA235564, R01CA195524, and P30CA016672. We'd like to thank Dr. Eugene J. Koay, Ph.D., for providing the liver MRI data from MD Anderson and Sarah Bronson - Scientific Editor at the Research Medical Library at MD Anderson - for editing sections of this article.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding the treatment of human subjects. This retrospective study was approved by The University of Texas MD Anderson Cancer Center Institutional Review Board (PA18-0832). The requirement for written informed consent was waived for this retrospective analysis.

Conflicts of Interest

We declare we have no conflicts of interest.

Data Availability

The DLDS dataset is available at `https://zenodo.org/records/7774566`. The AMOS dataset can be accessed at `https://zenodo.org/records/7155725`. The CHAOS dataset can be downloaded from `https://zenodo.org/records/3431873`. The ATLAS dataset is hosted at `https://atlas-challenge.u-bourgogne.fr/`. The BTCV challenge dataset is available at https://www.synapse.org/Synapse:syn3193805/wiki/217789. The MD Anderson data may be made available upon request in compliance with institutional IRB requirements.

Code Availability

Our code can be found at https://github.com/mckellwoodland/dimen_reduce_mahal (Woodland et al., 2024).

References

Jadie Adams and Shireen Y. Elhabian. Benchmarking scalable epistemic uncertainty quantification in organ segmentation. In Carole H. Sudre, Christian F. Baumgartner, Adrian Dalca, Raghav Mehta, Chen Qin, and William M. Wells, editors, *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 53–63, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-44336-7.

Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional space. In Jan Van den Bussche and Victor Vianu, editors, *ICDT 2001*, pages 420–434, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. ISBN 978-3-540-44503-6.

Brian M. Anderson, Ethan Y. Lin, Carlos E. Cardenas, Dustin A. Gress, William D. Erwin, Bruno C. Odisio, Eugene J. Koay, and Kristy K. Brock. Automated contouring of contrast and noncontrast computed tomography liver images with fully convolutional networks. *Adv. Radiat. Oncol*, 6(1):100464, 2021. ISSN 2452-1094. .

Harry Anthony and Konstantinos Kamnitsas. On the use of mahalanobis distance for out-of-distribution detection with neural networks for medical imaging. In Carole H. Sudre, Christian F. Baumgartner, Adrian Dalca, Raghav Mehta, Chen Qin, and William M. Wells, editors, *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 136–146, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-44336-7.

Hana Baroudi, Kristy K. Brock, Wenhua Cao, Xinru Chen, Caroline Chung, Laurence E. Court, Mohammad D. El Basha, Maguy Farhat, Skylar Gay, Mary P. Gronberg, Aashish Chandra Gupta, Soleil Hernandez, Kai Huang, David A. Jaffray, Rebecca Lim, Barbara Marquez, Kelly Nealon, Tucker J. Netherton, Callistus M. Nguyen, Brandon Reber, Dong Joo Rhee, Ramon M. Salazar, Mihir D. Shanker, Carlos Sjogreen, McKell Woodland, Jinzhong Yang, Cenji Yu, and Yao Zhao. Automated contouring and planning in radiation therapy: What is 'clinically acceptable'? *Diagnostics*, 13(4), 2023. ISSN 2075-4418. . URL https://www.mdpi.com/2075-4418/13/4/667.

Carlos E. Cardenas, Jinzhong Yang, Brian M. Anderson, Laurence E. Court, and Kristy B. Brock. Advances in auto-segmentation. *Seminars in Radiation Oncology*, 29(3):185–197, 2019. ISSN 1053-4296. . URL https://www.sciencedirect.com/science/article/pii/S1053429619300104. Adaptive Radiotherapy and Automation.

Wenqi Chen, Chi-Leung Chiang, and Laura A. Dawson. Efficacy and safety of radiotherapy for primary liver cancer. *Chinese Clinical Oncology*, 10(1), 2020. ISSN 2304-3873. URL `https://cco.amegroups.org/article/view/45511`.

Zheng Chen, Will King, Robert Pearcey, Marc Kerba, and William J. Mackillop. The relationship between waiting time for radiotherapy and clinical outcomes: A systematic review of the literature. *Radiotherapy and Oncology*, 87(1):3 – 16, 2008. ISSN 0167-8140. . URL `http://www.sciencedirect.com/science/article/pii/S0167814007005889`.

MONAI Consortium. Monai: Medical open network for ai, November 2021.

Jacques Ferlay, Murielle Colombet, Isabelle Soerjomataram, Donald M. Parkin, Marion Piñeros, Ariana Znaor, and Freddie Bray. Cancer statistics for the year 2020: An overview. *International Journal of Cancer*, 149(4):778–789, Apr 2021. .

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Adv. Neural Inf. Process. Syst.*, volume 34, pages 7068–7081. Curran Associates, Inc., 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *PMLR*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR.

Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/84ddfb34126fc3a48ee38d7044e87276-Paper.pdf`.

Soumya Suvra Ghosal, Yiyou Sun, and Yixuan Li. How to overcome curse-of-dimensionality for out-of-distribution detection? *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):19849–19857, Mar. 2024. . URL `https://ojs.aaai.org/index.php/AAAI/article/view/29960`.

Camila González, Karol Gotkowski, Andreas Bucher, Ricarda Fischbach, Isabel Kaltenborn, and Anirban Mukhopadhyay. Detecting when pre-trained nnu-net models fail silently for covid-19 lung lesion segmentation. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *MICCAI 2021*, pages 304–314, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87234-2. .

Camila González, Karol Gotkowski, Moritz Fuchs, Andreas Bucher, Armin Dadras, Ricarda Fischbach, Isabel Jasmin Kaltenborn, and Anirban Mukhopadhyay. Distance-based detection of out-of-distribution silent failures for covid-19 lung lesion segmentation. *Medical Image Analysis*, 82:102596, 2022. ISSN 1361-8415. . URL `https://www.sciencedirect.com/science/article/pii/S1361841522002298`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *ICML 2017*, volume 70 of *PMLR*, pages 1321–1330. PMLR, 06–11 Aug 2017.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks, 2017.

Fabian Isensee, Paul F. Jaeger, Simon A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. Nnu-net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, Dec 2020. .

Yuanfeng Ji. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation, November 2022.

Yuanfeng Ji, Haotian Bai, Chongjian GE, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, and Ping Luo. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Adv. Neural Inf. Process. Syst.*, volume 35, pages 36722–36732. Curran Associates, Inc., 2022.

Jue Jiang, Neelam Tyagi, Kathryn Tringale, Christopher Crane, and Harini Veeraraghavan. Self-supervised 3d anatomy segmentation using self-distilled masked image transformer (smit). In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *MICCAI 2022*, pages 556–566, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-16440-8.

Alain Jungo, Fabian Balsiger, and Mauricio Reyes. Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. *Frontiers in Neuroscience*, 14, 2020. ISSN 1662-453X. . URL `https://www.frontiersin.org/articles/10.3389/fnins.2020.00282`.

Davood Karimi and Ali Gholipour. Improving calibration and out-of-distribution detection in deep models for medical image segmentation. *IEEE Transactions on Artificial Intelligence*, 4(2):383–397, 2023. .

A. Emre Kavur, Naciye Sinem Gezer, Mustafa Barış, Yusuf Şahin, Savaş Özkan, Bora Baydar, Ulaş Yüksel, Çağlar Kılıkçıer, Şahin Olut, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. Comparison of semi-automatic and deep learning based automatic methods for liver segmentation in living liver transplant donors. *Diagnostic and Interventional Radiology*, 26:11–21, January 2020. . URL `https://doi.org/10.5152/dir.2019.19025`.

A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, Bora Baydar, Dmitry Lachinov, Shuo Han, Josef Pauli, Fabian Isensee, Matthias Perkonigg, Rachana Sathish, Ronnie Rajan, Debdoot Sheet, Gurbandurdy Dovletov, Oliver Speck, Andreas Nürnberger, Klaus H. Maier-Hein, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. Chaos challenge - combined (ct-mr) healthy abdominal organ segmentation. *Med. Image Anal.*, 69:101950, 2021. ISSN 1361-8415. .

Ali Emre Kavur, M. Alper Selver, Oğuz Dicle, Mustafa Barış, and N. Sinem Gezer. CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data, April 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Adv. Neural Inf. Process. Syst.*, volume 30. Curran Associates, Inc., 2017.

Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, page 12, 2015.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Adv. Neural Inf. Process. Syst.*, volume 31. Curran Associates, Inc., 2018.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks, 2018.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Adv. Neural Inf. Process. Syst.*, volume 33, pages 21464–21475. Curran Associates, Inc., 2020.

Jacob A. Macdonald, Zhe Zhu, Brandon Konkel, Maciej Mazurowski, Walter Wiggins, and Mustafa Bashir. Duke liver dataset (mri), October 2020.

Jacob A. Macdonald, Zhe Zhu, Brandon Konkel, Maciej A. Mazurowski, Walter F. Wiggins, and Mustafa R. Bashir. Duke liver dataset: A publicly available liver mri dataset with liver segmentation masks and series labels. *Radiology: Artificial Intelligence*, 5(5): e220275, 2023. . URL `https://doi.org/10.1148/ryai.220275`. PMID: 37293348.

P. C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55, 1936.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

Alireza Mehrtash, William M. Wells, Clare M. Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, 39(12):3868–3878, 2020. .

Multi-Institutional. Human-computer interaction in radiotherapy target volume delineation: A prospective, multi-institutional comparison of user input devices. *Journal of Digital Imaging*, 24(5):794–803, Oct 2011. .

Benjamin E Nelms, Wolfgang A Tomé, Greg Robinson, and James Wheeler. Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer. *International Journal in Oncology Biology Physics*, 82(1):368–378, Jan 2012. .

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Ricardo Otazo, Philippe Lambin, Jean-Philippe Pignol, Mark E. Ladd, Heinz-Peter Schlemmer, Michael Baumann, and Hedvig Hricak. Mri-guided radiation therapy: An emerging paradigm in adaptive radiation oncology. *Radiology*, 298(2):248–260, 2021. . URL `https://doi.org/10.1148/radiol.2020202747`. PMID: 33350894.

Nihil Patel, Mohamed Eltaher, Rachel Glenn, Kari Brewer Savannah, Kristy Brock, Jessica Sanchez, Tiffany Calderone, Darrel Cleere, Ahmed Elsaiey, Matthew Cagley, Nakul Gupta, David Victor, Laura Beretta, Adrian Celaya, Eugene Koay, Tucker Netherton, and David Fuentes. Training robust t1-weighted magnetic resonance imaging liver segmentation models using ensembles of datasets with different contrast protocols and liver disease etiologies. 2024. .

Geoff Pleiss, Amauri Souza, Joseph Kim, Boyi Li, and Kilian Q Weinberger. Neural network out-of-distribution detection for regression tasks. 2019.

Félix Quinton, Romain Popoff, Benoît Presles, Sarah Leclerc, Fabrice Meriaudeau, Guillaume Nodari, Olivier Lopez, Julie Pellegrinelli, Olivier Chevallier, Dominique Ginhac, Jean-Marc Vrigneaud, and Jean-Louis Alberini. A tumour and liver automatic segmentation (atlas) dataset on contrast-enhanced magnetic resonance imaging for hepatocellular carcinoma. *Data*, 8(5), 2023. ISSN 2306-5729. . URL `https://www.mdpi.com/2306-5729/8/5/79`.

A. Blythe Ryerson, Christie R. Eheman, Sean F. Altekruse, John W. Ward, Ahmedin Jemal, Recinda L. Sherman, S. Jane Henley, Deborah Holtzman, Andrew Lake, Anne-Michelle Noone, and et al. Annual report to the nation on the status of cancer, 1975-2012, featuring the increasing incidence of liver cancer. *Cancer*, 122(9):1312–1337, Mar 2016. .

Anne E. Saarnak, Menno Boersma, Bart N.F.M. van Bunningen, René Wolterink, and Marcel J. Steggerda. Inter-observer variation in delineation of bladder and rectum contours for brachytherapy of cervical cancer. *Radiotherapy and Oncology*, 56(1):37 – 42, 2000. ISSN 0167-8140. . URL `http://www.sciencedirect.com/science/article/pii/S0167814000001857`.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with Gram matrices. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8491–8501. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/sastry20a.html`.

Ke Sheng. Artificial intelligence in radiotherapy: a technological review. *Frontiers of Medicine*, 14(4):431, 2020. . URL `http://journal.hep.com.cn/fmd/EN/abstract/article_27599.shtml`.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Adv. Neural Inf. Process. Syst.*, volume 34, pages 144–157. Curran Associates, Inc., 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20827–20840. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/sun22d.html`.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R. Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *CVPR 2022*, pages 20730–20740, June 2022.

Mattias Teye, Hossein Azizpour, and Kevin Smith. Bayesian uncertainty estimation for batch normalized deep networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4907–4916. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/teye18a.html`.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *J. Mach. Learn. Res.*, 9(11), 2008.

Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: An alternative approach to efficient ensemble and lifelong learning, 2020.

McKell Woodland, Nihil Patel, Mais Al Taie, Joshua P. Yung, Tucker J. Netherton, Ankit B. Patel, and Kristy K. Brock. Dimensionality reduction for improving out-of-distribution detection in medical image segmentation. In Carole H. Sudre, Christian F. Baumgartner, Adrian Dalca, Raghav Mehta, Chen Qin, and William M. Wells, editors, *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 147–156, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-44336-7.

McKell Woodland, Ankit B. Patel, and Kristy K. Brock. Dimensionality Reduction and Nearest Neighbors for Improving Out-of-Distribution Detection in Medical Image Segmentation - Official Repository, October 2024. URL `https://doi.org/10.5281/zenodo.13881989`.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey, 2024.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11): 1–17, 11 2018. .

Appendix A. OOD detection results split by threshold

A.1 Mahalanobis Distance

Table 8: MD-based OOD detection of poor performance results. ID is a DSC $\geq$ 95%. Only the best-performing configuration by AUROC is reported for each dimensionality reduction technique (Reduct): no reduction (None), PCA(n_p) with n_p components, t-SNE, UMAP(n_u) with n_u components, and n_d-dimensional average pooling with kernel size k and stride s, Pooln_dD(k,s). Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether higher or lower is better. Bold highlights the best performance per model.

Model	**Reduct**	**AUROC ↑**	**AUPRC ↑**	**FPR90 ↓**	**Seconds ↓**
MRI UNETR	None	0.48 (±0.00)	0.61 (±0.00)	1.00 (±0.00)	9,354.34 (±48.53)
	PCA(256)	**0.93** (±0.00)	**0.94** (±0.00)	**0.23** (±0.00)	**2.82** (±0.14)
	t-SNE	0.70 (±0.08)	0.72 (±0.12)	0.71 (±0.14)	4.70 (±0.28)
	UMAP(2)	0.77 (±0.08)	0.79 (±0.11)	0.57 (±0.08)	10.44 (±0.22)
	Pool2D(3,2)	0.82 (±0.00)	0.86 (±0.00)	0.46 (±0.00)	15.32 (±8.92)
MRI+ UNETR	None	0.46 (±0.00)	**1.00** (±0.00)	1.00 (±0.00)	9809.40 (±57.20)
	PCA(16)	0.82 (±0.00)	**1.00** (±0.00)	1.00 (±0.00)	**1.42** (±0.06)
	t-SNE	0.92 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	5.75 (±0.16)
	UMAP(16)	0.91 (±0.03)	**1.00** (±0.00)	0.20 (±0.40)	16.69 (0.98)
	Pool3D(3,1)	**0.96** (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	78.77 (±0.20)
MRI+ nnU-net	None	0.69 (±0.00)	**1.00** (±0.00)	0.67 (±0.00)	4,125.96 (±13.12)
	PCA(8)	**0.96** (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	**1.12** (±0.04)
	t-SNE	0.70 (±0.18)	**1.00** (±0.00)	0.87 (±0.16)	4.72 (±0.11)
	UMAP(16)	0.82 (±0.08)	**1.00** (±0.00)	0.67 (±0.03)	19.07 (±0.96)
	Pool2D(2,1)	0.85 (±0.00)	**1.00** (±0.00)	0.67 (±0.00)	1,579.73 (±52.41)
CT nnU-net	None	0.41 (±0.00)	0.10 (±0.00)	0.91 (±0.00)	5,856.21 (±63.04)
	PCA(32)	0.56 (±0.00)	0.17 (±0.00)	0.92 (±0.00)	**8.17** (±0.23)
	t-SNE	0.59 (±0.04)	0.20 (±0.02)	0.80 (±0.02)	13.75 (±0.35)
	UMAP(128)	**0.68** (±0.03)	**0.22** (±0.01)	**0.74** (±0.09)	288.42 (±25.99)
	Pool2D(2,2)	0.59 (±0.00)	0.13 (±0.00)	0.84 (±0.00)	163.84 (±19.54)

Table 9: MD-based OOD detection of poor performance results. ID is a DSC ≥ 80%. Only the best-performing configuration by AUROC is reported for each dimensionality reduction technique (Reduct): no reduction (None), PCA(n_p) with n_p components, t-SNE, UMAP(n_u) with n_u components, and n_d-dimensional average pooling with kernel size k and stride s, Pooln_dD(k,s). Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether higher or lower is better. Bold highlights the best performance per model.

Model	**Reduct**	**AUROC ↑**	**AUPRC ↑**	**FPR90 ↓**	**Seconds ↓**
MRI UNETR	None	0.48 (±0.00)	0.21 (±0.00)	0.83 (±0.00)	9349.73 (±10.85)
	PCA(128)	0.92 (±0.00)	0.68 (±0.00)	**0.13** (±0.00)	1.95 (±0.12)
	t-SNE	0.85 (±0.00)	0.49 (±0.00)	0.26 (±0.00)	4.35 (±0.13)
	UMAP(8)	**0.93** (±0.03)	**0.73** (±0.07)	**0.13** (±0.00)	5.61 (±0.14)
	Pool3D(4,1)	0.87 (±0.00)	0.50 (±0.00)	0.22 (±0.00)	**1.72** (±0.06)
MRI+ UNETR	None	0.53 (±0.00)	0.03 (±0.00)	0.78 (±0.00)	10,070.78 (±141.69)
	PCA(32)	**0.85** (±0.01)	**0.13** (±0.00)	**0.34** (±0.02)	**1.86** (±0.32)
	t-SNE	0.66 (±0.00)	0.03 (±0.00)	0.45 (±0.00)	5.77 (±0.06)
	UMAP(2)	0.68 (±0.07)	0.05 (±0.01)	0.49 (±0.06)	21.29 (±0.43)
	Pool2D(4,1)	0.64 (±0.00)	0.04 (±0.00)	0.64 (±0.00)	16.08 (±13.41)
MRI+ nnU-net	None	0.58 (±0.00)	0.03 (±0.00)	0.70 (±0.00)	4,100.82 (±4.00)
	PCA(8)	0.74 (±0.00)	0.08 (±0.00)	0.52 (±0.00)	**1.17** (±0.04)
	t-SNE	0.35 (±0.00)	0.02 (±0.00)	0.90 (±0.00)	4.68 (±0.06)
	UMAP(4)	**0.78** (±0.00)	0.06 (±0.02)	**0.45** (±0.07)	13.43 (±0.24)
	Pool3D(2,2)	0.68 (±0.00)	**0.30** (±0.00)	0.71 (±0.00)	6.99 (±0.18)
CT nnU-net	None	-	-	-	-
	PCA	-	-	-	-
	t-SNE	-	-	-	-
	UMAP	-	-	-	-
	Pool	-	-	-	-

Table 10: MD-based OOD detection of poor performance results. ID is a DSC $\geq$ the median DSC. Only the best-performing configuration by AUROC is reported for each dimensionality reduction technique (Reduct): no reduction (None), PCA(n_p) with n_p components, t-SNE, UMAP(n_u) with n_u components, and n_d-dimensional average pooling with kernel size k and stride s, Pooln_dD(k,s). Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether higher or lower is better. Bold highlights the best performance per model.

Model	**Reduct**	**AUROC ↑**	**AUPRC ↑**	**FPR90 ↓**	**Seconds ↓**
MRI UNETR	None	0.48 (±0.00)	0.61 (±0.00)	1.00 (±0.00)	9,283.76 (±8.18)
	PCA(256)	**0.93** (±0.00)	**0.94** (±0.00)	**0.23** (±0.00)	**2.93** (±0.14)
	t-SNE	0.60 (±0.00)	0.57 (±0.00)	0.77 (±0.00)	4.42 (±0.09)
	UMAP(8)	0.74 (±0.05)	0.83 (±0.04)	0.74 (±0.10)	5.79 (±0.22)
	Pool2D(3,2)	0.82 (±0.00)	0.86 (±0.00)	0.46 (±0.00)	4.70 (±0.14)
MRI+ UNETR	None	0.50 (±0.00)	0.52 (±0.00)	0.90 (±0.00)	9,777.69 (±12.41)
	PCA(32)	0.54 (±0.00)	0.58 (±0.00)	0.86 (±0.00)	**1.66** (±0.15)
	t-SNE	0.52 (±0.00)	0.54 (±0.00)	0.90 (±0.00)	5.73 (±0.16)
	UMAP(32)	0.55 (±0.00)	0.55 (±0.01)	**0.84** (±0.02)	17.01 (±0.57)
	Pool2D(4,1)	**0.57** (±0.00)	**0.61** (±0.00)	**0.84** (±0.00)	9.38 (±0.14)
MRI+ nnU-net	None	0.51 (±0.00)	0.51 (±0.00)	0.88 (±0.00)	4,091.49 (±5.61)
	PCA(32)	**0.62** (±0.00)	**0.62** (±0.00)	**0.80** (±0.00)	**1.39** (±0.09)
	t-SNE	0.59 (±0.00)	0.60 (±0.00)	0.82 (±0.00)	4.65 (±0.07)
	UMAP(64)	0.58 (±0.02)	0.58 (±0.01)	0.86 (±0.03)	14.16 (±0.16)
	Pool3D(2,2)	0.57 (±0.00)	0.58 (±0.00)	0.86 (±0.00)	6.86 (±0.11)
CT nnU-net	None	0.49 (±0.00)	0.49 (±0.00)	0.90 (±0.00)	5,388.67 (±47.96)
	PCA(2)	0.56 (±0.00)	0.56 (±0.00)	0.88 (±0.00)	**4.79** (±0.06)
	t-SNE	**0.63** (±0.00)	**0.66** (±0.00)	**0.85** (±0.00)	12.09 (±0.17)
	UMAP(32)	0.57 (±0.03)	0.58 (±0.03)	**0.85** (±0.04)	199.15 (±1.56)
	Pool2D(4,1)	0.57 (±0.00)	0.58 (±0.00)	**0.85** (±0.00)	130.59 (±1.80)

A.2 K-Nearest Neighbors

Table 11: KNN-based OOD detection of poor performance results. ID is a DSC ≥ 95%. Only the best-performing configuration by AUROC is reported for each dimensionality reduction technique (Reduct): no reduction (None), PCA(n_p) with n_p components, t-SNE, UMAP(n_u) with n_u components, and n_d-dimensional average pooling with kernel size k and stride s, Pooln_dD(k,s). Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether higher or lower is better. Bold highlights the best performance per model.

Model	**Reduction**	**K**	**AUROC ↑**	**AUPRC ↑**	**FPR90 ↓**	**Seconds ↓**
MRI UNETR	None	256	0.87 (±0.00)	0.88 (±0.00)	0.31 (±0.00)	**0.78** (±0.00)
	PCA(2)	256	0.90 (±0.00)	0.92 (±0.00)	0.31 (±0.00)	0.95 (±0.00)
	t-SNE	256	0.77 (±0.05)	0.83 (±0.04)	0.74 (±0.06)	4.48 (±0.07)
	UMAP(32)	256	0.83 (±0.05)	0.85 (±0.04)	0.51 (±0.08)	6.70 (±0.15)
	Pool2D(3,1)	256	**0.94** (±0.00)	**0.95** (±0.00)	**0.23** (±0.00)	0.90 (±0.00)
MRI+ UNETR	None	32	0.86 (±0.00)	**1.00** (±0.00)	1.00 (±0.00)	9.27 (±0.15)
	PCA(16)	8	0.88 (±0.02)	**1.00** (±0.00)	0.40 (±0.49)	1.52 (±0.06)
	t-SNE	256	0.94 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	5.67 (±0.08)
	UMAP(2)	64	0.89 (±0.02)	**1.00** (±0.00)	0.60 (±0.49)	16.47 (±0.28)
	Pool2D(3,2)	4	**0.97** (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	**1.30** (±0.03)
MRI+ nnU-net	None	256	0.96 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	6.78 (±0.10)
	PCA(8)	256	0.97 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	1.16 (±0.07)
	t-SNE	128	0.67 (±0.08)	**1.00** (±0.00)	1.00 (±0.00)	4.76 (±0.11)
	UMAP(2)	2	0.96 (±0.04)	**1.00** (±0.00)	0.80 (±0.27)	14.18 (±0.55)
	Pool2D(2,2)	256	**0.98** (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	**0.74** (±0.04)
CT nnU-net	None	8	0.52 (±0.00)	0.13 (±0.00)	0.94 (±0.00)	37.64 (±0.67)
	PCA(8)	4	0.55 (±0.00)	0.15 (±0.00)	0.97 (±0.00)	**4.94** (±0.04)
	t-SNE	256	0.46 (±0.00)	0.19 (±0.00)	0.95 (±0.01)	12.28 (±0.09)
	UMAP(4)	256	**0.65** (±0.01)	**0.24** (±0.05)	**0.88** (±0.02)	199.93 (±1.96)
	Pool3D(2,2)	4	0.54 (±0.00)	0.14 (±0.00)	0.96 (±0.00)	81.96 (±26.79)

Table 12: KNN-based OOD detection of poor performance results. ID is a DSC $\geq$ 80%. Only the best-performing configuration by AUROC is reported for each dimensionality reduction technique (Reduct): no reduction (None), PCA(n_p) with n_p components, t-SNE, UMAP(n_u) with n_u components, and n_d-dimensional average pooling with kernel size k and stride s, Pooln_dD(k,s). Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether higher or lower is better. Bold highlights the best performance per model.

Model	**Reduct**	**K**	**AUROC ↑**	**AUPRC ↑**	**FPR90 ↓**	**Seconds ↓**
MRI UNETR	None	256	0.87 (±0.00)	0.62 (±0.00)	**0.13** (±0.00)	**0.79** (±0.02)
	PCA(2)	128	**0.92** (±0.00)	0.67 (±0.00)	**0.13** (±0.00)	0.93 (±0.05)
	t-SNE	128	0.83 (±0.09)	0.44 (±0.11)	0.31 (±0.17)	4.36 (±0.07)
	UMAP(2)	256	0.90 (±0.05)	0.57 (±0.18)	0.16 (±0.08)	5.68 (±0.10)
	Pool2D(3,2)	256	**0.92** (±0.00)	**0.69** (±0.00)	**0.13** (±0.00)	0.84 (±0.08)
MRI+ UNETR	None	256	0.76 (±0.00)	0.25 (±0.00)	0.59 (±0.00)	9.44 (±0.16)
	PCA(32)	64	0.84 (±0.01)	0.17 (±0.02)	0.40 (±0.01)	**1.57** (±0.06)
	t-SNE	64	0.70 (±0.00)	0.04 (±0.00)	**0.37** (±0.00)	6.49 (±0.19)
	UMAP(4)	64	0.78 (±0.05)	0.09 (±0.06)	0.42 (±0.07)	15.70 (±0.45)
	Pool3D(2,2)	256	**0.87** (±0.00)	**0.33** (±0.00)	0.43 (±0.00)	11.21 (±7.66)
MRI+ nnU-net	None	256	0.72 (±0.00)	**0.09** (±0.00)	0.67 (±0.00)	6.44 (±0.09)
	PCA(8)	256	0.81 (±0.00)	0.06 (±0.00)	0.36 (±0.00)	1.12 (±0.07)
	t-SNE	2	0.74 (±0.10)	0.05 (±0.02)	0.44 (±0.20)	4.80 (±0.12)
	UMAP(128)	64	**0.82** (±0.00)	0.06 (±0.01)	0.39 (±0.02)	14.61 (±0.25)
	Pool3D(4,1)	64	0.76 (±0.00)	0.08 (±0.00)	**0.34** (±0.00)	**0.77** (±0.06)
CT nnU-net	None	-	-	-	-	-
	PCA	-	-	-	-	-
	t-SNE	-	-	-	-	-
	UMAP	-	-	-	-	-
	Pool	-	-	-	-	-

Table 13: KNN-based OOD detection of poor performance results. ID is a DSC $\geq$ the median DSC. Only the best-performing configuration by AUROC is reported for each dimensionality reduction technique (Reduct): no reduction (None), PCA(n_p) with n_p components, t-SNE, UMAP(n_u) with n_u components, and n_d-dimensional average pooling with kernel size k and stride s, Pooln_dD(k,s). Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether higher or lower is better. Bold highlights the best performance per model.

Model	**Reduct**	**K**	**AUROC ↑**	**AUPRC ↑**	**FPR90 ↓**	**Seconds ↓**
MRI UNETR	None	256	0.87 (±0.00)	0.88 (±0.00)	0.31 (±0.00)	**0.78** (±0.00)
	PCA(2)	256	0.90 (±0.00)	0.92 (±0.00)	0.31 (±0.00)	0.98 (±0.08)
	t-SNE	256	0.77 (±0.05)	0.83 (±0.04)	0.74 (±0.06)	4.37 (±0.11)
	UMAP(128)	256	0.82 (±0.04)	0.86 (±0.04)	0.63 (±0.16)	6.61 (±0.43)
	Pool2D(3,1)	256	**0.94** (±0.00)	**0.95** (±0.00)	**0.23** (±0.00)	0.89 (±0.01)
MRI+ UNETR	None	32	0.86 (±0.00)	**1.00** (±0.00)	1.00 (±0.00)	9.01 (±0.07)
	PCA(16)	8	0.88 (±0.02)	**1.00** (±0.00)	0.40 (±0.49)	**1.46** (±0.09)
	t-SNE	256	0.94 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	5.67 (±0.04)
	UMAP(2)	64	0.90 (±0.02)	**1.00** (±0.00)	0.40 (±0.49)	16.75 (±0.26)
	Pool2D(3,2)	4	**0.97** (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	1.29 (±0.01)
MRI+ nnU-net	None	256	0.96 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	6.76 (±0.00)
	PCA(8)	256	0.97 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	1.12 (±0.04)
	t-SNE	128	0.67 (±0.09)	**1.00** (±0.00)	1.00 (±0.00)	4.73 (±0.11)
	UMAP(2)	2	0.96 (±0.02)	**1.00** (±0.00)	0.20 (±0.16)	14.82 (±1.60)
	Pool2D(2,2)	256	**0.98** (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	**0.77** (±0.03)
CT nnU-net	None	8	0.52 (±0.00)	0.13 (±0.00)	0.94 (±0.00)	37.15 (±0.28)
	PCA(8)	4	0.55 (±0.00)	0.15 (±0.00)	0.97 (±0.00)	**5.03** (±0.08)
	t-SNE	256	0.46 (±0.00)	0.19 (±0.00)	0.95 (±0.01)	12.07 (±0.12)
	UMAP(2)	256	**0.66** (±0.01)	**0.23** (±0.05)	**0.88** (±0.02)	206.25 (±1.34)
	Pool3D(2,1)	4	0.54 (±0.00)	0.14 (±0.00)	0.96 (±0.00)	23.84 (±0.15)

A.3 Comparison Methods

Table 14: OOD detection results for the best-performing configurations of the comparison methods: MSP, temperature scaling (TS) and energy scoring (Energy) with temperature T, MC Dropout (MCD), and ensembling (Ensemble). ID is a DSC $\geq$ 95%. Seconds is the amount of time it took to calculate the test distances. Only the best-performing configuration by AUROC is reported for each method. The results are averages (±SD) across 5 runs. Arrows denote whether a higher or lower value is better. Bold highlights the best performance per model.

Model	**Method**	**T**	**AUROC ↑**	**AUPRC ↑**	**FPR90 ↓**	**Seconds ↓**
MRI UNETR	MSP	-	**0.96** (±0.00)	**0.97** (±0.00)	**0.00** (±0.00)	8.78 (±0.11)
	TS	2	0.90 (±0.00)	0.93 (±0.00)	0.15 (±0.00)	9.25 (±0.11)
	Energy	3	0.55 (±0.00)	0.57 (±0.00)	0.69 (±0.00)	**7.08** (±0.14)
Dropout UNETR	MCD	-	**1.00** (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	**14.56** (±0.20)
Ensemble UNETR	Ensemble	-	**0.96** (±0.00)	**0.96** (±0.00)	**0.14** (±0.00)	**14.02** (±0.05)
MRI+ UNETR	MSP	-	0.91 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	59.18 (±0.46)
	TS	2	**0.93** (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	60.57 (±0.43)
	Energy	1	0.89 (±0.00)	**1.00** (±0.00)	1.00 (±0.00)	**36.53** (±1.82)
MRI+ nnU-net	MSP	-	0.45 (±0.00)	0.99 (±0.00)	**1.00** (±0.00)	1,114.34 (±0.51)
	TS	10	0.55 (±0.00)	0.99 (±0.00)	**1.00** (±0.00)	1,252.78 (±0.73)
	Energy	10	**0.61** (±0.00)	**1.00** (±0.00)	**1.00** (±0.00)	**186.88** (±0.47)
CT nnU-net	MSP	-	**0.72** (±0.00)	**0.29** (±0.00)	**0.57** (±0.00)	568.51 (±0.37)
	TS	3	0.68 (±0.00)	0.26 (±0.00)	0.69 (±0.00)	699.07 (±0.37)
	Energy	2	0.67 (±0.00)	0.26 (±0.00)	0.75 (±0.00)	**105.66** (±0.60)

Table 15: OOD detection results for the best-performing configurations of the comparison methods: MSP, temperature scaling (TS) and energy scoring (Energy) with temperature T, MC Dropout (MCD), and ensembling (Ensemble). ID is a DSC $\geq$ 80%. Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether a higher or lower value is better. Bold highlights the best performance per model.

Model	**Method**	**T**	**AUROC ↑**	**AUPRC ↑**	**FPR90 ↓**	**Seconds ↓**
MRI UNETR	MSP	-	**0.92** (±0.00)	**0.58** (±0.00)	**0.09** (±0.00)	8.84 (±0.14)
	TS	2	0.91 (±0.00)	0.54 (±0.00)	**0.09** (±0.00)	9.10 (±0.14)
	Energy	1000	0.72 (±0.00)	0.26 (±0.00)	0.35 (±0.00)	**6.70** (±0.18)
MRI Dropout	MCD	-	**1.00** (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	**14.70** (±0.23)
MRI Ensemble	Ensemble	-	**0.96** (±0.00)	**0.83** (±0.00)	**0.08** (±0.00)	**14.16** (±0.20)
MRI+ UNETR	MSP	-	0.57 (±0.00)	**0.09** (±0.00)	**0.59** (±0.00)	58.70 (±0.48)
	TS	2	0.47 (±0.00)	0.05 (±0.00)	0.78 (±0.00)	60.57 (±0.31)
	Energy	1000	**0.61** (±0.00)	0.03 (±0.00)	0.71 (±0.00)	**35.86** (±0.41)
MRI+ nnU-net	MSP	-	0.87 (±0.00)	0.10 (±0.00)	0.22 (±0.00)	1,115.75 (±0.60)
	TS	3	**0.92** (±0.00)	**0.15** (±0.00)	**0.14** (±0.00)	1,252.07 (±0.22)
	Energy	1	0.68 (±0.00)	0.09 (±0.00)	0.75 (±0.00)	**185.96** (±0.39)
CT nnU-net	MSP	-	-	-	-	-
	TS	-	-	-	-	-
	Energy	-	-	-	-	-

Table 16: OOD detection results for the best-performing configurations of the comparison methods: MSP, temperature scaling (TS) and energy scoring (Energy) with temperature T, MC Dropout (MCD), and ensembling (Ensemble). ID is a DSC $\geq$ the median DSC. Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether a higher or lower value is better. Bold highlights the best performance per model.

Model	**Method**	**T**	**AUROC ↑**	**AUPRC ↑**	**FPR90 ↓**	**Seconds ↓**
MRI UNETR	MSP	-	**0.96** (±0.00)	**0.97** (±0.00)	**0.00** (±0.00)	8.73 (±0.12)
	TS	2	0.90 (±0.00)	0.93 (±0.00)	0.15 (±0.00)	9.14 (±0.05)
	Energy	4	0.64 (±0.00)	0.56 (±0.00)	0.60 (±0.00)	**6.95** (±0.26)
MRI Dropout	MCD	-	**1.00** (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	**14.57** (±0.14)
MRI Ensemble	Ensemble	-	**0.96** (±0.00)	**0.96** (±0.00)	**0.14** (±0.00)	**13.93** (±0.09)
MRI+ UNETR	MSP	-	0.60 (±0.00)	0.62 (±0.00)	**0.85** (±0.00)	58.31 (±0.21)
	TS	5	**0.63** (±0.00)	**0.65** (±0.00)	**0.85** (±0.00)	60.77 (±0.29)
	Energy	2	0.62 (±0.00)	0.63 (±0.00)	**0.85** (±0.00)	**42.34** (±0.26)
MRI+ nnU-net	MSP	-	0.51 (±0.00)	0.56 (±0.00)	0.99 (±0.00)	1,114.55 (±0.17)
	TS	5	**0.64** (±0.00)	**0.66** (±0.00)	0.85 (±0.00)	1,252.77 (±0.19)
	Energy	1	0.63 (±0.00)	0.64 (±0.00)	**0.78** (±0.00)	**190.26** (±0.42)
CT nnU-net	MSP	-	**0.67** (±0.00)	**0.71** (±0.00)	**0.91** (±0.00)	567.86 (±0.33)
	TS	3	0.66 (±0.00)	0.70 (±0.00)	**0.91** (±0.00)	699.59 (±0.24)
	Energy	1	0.65 (±0.00)	0.70 (±0.00)	0.93 (±0.00)	**100.77** (±0.84)

Appendix B. Hyperparameter searches

B.1 Mahalanobis Distance

Table 17: MD hyperparameter searches for the MRI UNETR (95% threshold). Dimensionality reduction techniques include PCA(n_p) with n_p components, UMAP(n_u) with n_u components, and n_d-dimensional average pooling with kernel size k and stride s, Pooln_dD(k,s). Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether higher or lower is better. Bold highlights the best performance per technique.

Experiment	**AUROC ↑**	**AUPRC ↑**	**FPR90 ↓**	**Seconds ↓**
PCA(2)	0.90 (±0.00)	0.93 (±0.00)	0.38 (±0.00)	0.95 (±0.06)
PCA(4)	0.70 (±0.00)	0.66 (±0.00)	0.46 (±0.00)	1.02 (±0.04)
PCA(8)	0.73 (±0.00)	0.74 (±0.00)	0.54 (±0.00)	1.02 (±0.10)
PCA(16)	0.87 (±0.00)	0.87 (±0.00)	**0.23** (±0.00)	1.13 (±0.11)
PCA(32)	0.85 (±0.01)	0.88 (±0.01)	0.34 (±0.08)	1.18 (±0.08)
PCA(64)	0.85 (±0.01)	0.87 (±0.02)	**0.23** (±0.00)	1.31 (±0.18)
PCA(128)	0.91 (±0.00)	0.93 (±0.00)	**0.23** (±0.00)	1.99 (±0.09)
PCA(256)	**0.93** (±0.00)	**0.94** (±0.00)	**0.23** (±0.00)	2.82 (±0.14)
UMAP(2)	**0.77** (±0.08)	0.79 (±0.11)	**0.57** (±0.08)	10.44 (±0.22)
UMAP(4)	0.75 (±0.04)	0.80 (±0.07)	0.66 (±0.16)	10.63 (±0.39)
UMAP(8)	0.74 (±0.05)	**0.82** (±0.03)	0.72 (±0.17)	10.57 (±0.16)
UMAP(16)	0.65 (±0.03)	0.77 (±0.02)	0.91 (±0.09)	10.76 (±0.29)
UMAP(32)	0.66 (±0.03)	0.77 (±0.02)	0.91 (±0.06)	10.64 (±0.38)
UMAP(64)	0.63 (±0.03)	0.75 (±0.03)	0.88 (±0.06)	10.76 (±0.33)
UMAP(128)	0.65 (±0.03)	0.77 (±0.01)	0.88 (±0.01)	10.99 (±0.45)
UMAP(256)	0.63 (±0.03)	0.76 (±0.02)	0.88 (±0.08)	11.17 (±0.28)
Pool2D(2,1)	0.71 (±0.00)	0.74 (±0.00)	0.54 (±0.00)	1,446.90 (±11.62)
Pool2D(2,2)	0.63 (±0.00)	0.69 (±0.00)	0.85 (±0.00)	187.00 (±10.68)
Pool2D(3,1)	0.72 (±0.00)	0.72 (±0.00)	0.54 (±0.00)	145.02 (±0.24)
Pool2D(3,2)	**0.82** (±0.00)	**0.86** (±0.00)	**0.46** (±0.00)	15.32 (±8.92)
Pool2D(4,1)	0.73 (±0.00)	0.78 (±0.00)	0.77 (±0.00)	11.95 (±5.92)
Pool3D(2,1)	0.70 (±0.00)	0.78 (±0.00)	0.92 (±0.00)	1,109.29 (±12.88)
Pool3D(2,2)	0.60 (±0.00)	0.69 (±0.00)	0.77 (±0.00)	18.69 (±0.25)
Pool3D(3,1)	0.75 (±0.00)	0.82 (±0.00)	0.85 (±0.00)	74.33 (±15.52)
Pool3D(3,2)	0.54 (±0.00)	0.58 (±0.00)	0.85 (±0.00)	1.10 (±0.04)
Pool3D(4,1)	0.70 (±0.00)	0.74 (±0.00)	0.54 (±0.00)	1.89 (±0.16)

Table 18: MD hyperparameter searches for the MRI+ UNETR (80% threshold). Dimensionality reduction techniques include PCA(n_p) with n_p components, UMAP(n_u) with n_u components, and n_d-dimensional average pooling with kernel size k and stride s, Pooln_dD(k,s). Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether higher or lower is better. Bold highlights the best performance per technique.

Experiment	**AUROC ↑**	**AUPRC ↑**	**FPR90 ↓**	**Seconds ↓**
PCA(2)	0.60 (±0.00)	0.03 (±0.00)	0.57 (±0.00)	1.59 (±0.03)
PCA(4)	0.57 (±0.00)	0.03 (±0.00)	0.62 (±0.00)	1.72 (±0.09)
PCA(8)	0.80 (±0.00)	**0.20** (±0.00)	0.46 (±0.00)	1.56 (±0.03)
PCA(16)	0.84 (±0.01)	0.10 (±0.01)	0.35 (±0.02)	1.60 (±0.23)
PCA(32)	**0.85** (±0.01)	0.13 (±0.00)	0.34 (±0.02)	1.86 (±0.32)
PCA(64)	0.80 (±0.01)	0.06 (±0.00)	0.36 (±0.02)	1.71 (±0.05)
PCA(128)	0.75 (±0.02)	0.04 (±0.00)	0.37 (±0.01)	2.88 (±0.13)
PCA(256)	0.75 (±0.01)	0.04 (±0.00)	**0.32** (±0.01)	3.57 (±0.10)
UMAP(2)	**0.68** (±0.07)	**0.05** (±0.01)	**0.49** (±0.06)	21.29 (±0.43)
UMAP(4)	0.64 (±0.07)	0.04 (±0.00)	0.50 (±0.07)	21.37 (±0.44)
UMAP(8)	0.63 (±0.02)	0.03 (±0.00)	0.51 (±0.05)	21.39 (±0.39)
UMAP(16)	0.60 (±0.04)	0.03 (±0.01)	0.54 (±0.05)	21.07 (±0.48)
UMAP(32)	0.57 (±0.04)	0.03 (±0.00)	0.66 (±0.03)	21.35 (±0.64)
UMAP(64)	0.53 (±0.03)	0.03 (±0.01)	0.64 (±0.05)	21.73 (±0.77)
UMAP(128)	0.54 (±0.04)	0.03 (±0.00)	0.61 (±0.04)	21.90 (±0.71)
UMAP(256)	0.55 (±0.04)	0.03 (±0.00)	0.60 (±0.03)	22.47 (±0.51)
Pool2D(2,1)	0.60 (±0.00)	0.03 (±0.00)	0.81 (±0.00)	1,602.81 (±48.97)
Pool2D(2,2)	0.62 (±0.00)	0.03 (±0.00)	0.63 (±0.00)	218.32 (±33.26)
Pool2D(3,1)	0.32 (±0.00)	0.02 (±0.00)	0.96 (±0.00)	208.96 (±24.75)
Pool2D(3,2)	0.58 (±0.00)	0.03 (±0.00)	0.73 (±0.00)	36.33 (±8.93)
Pool2D(4,1)	**0.64** (±0.00)	0.04 (±0.00)	0.64 (±0.00)	16.08 (±13.41)
Pool3D(2,1)	0.57 (±0.00)	0.02 (±0.00)	**0.61** (±0.00)	1,338.82 (±99.92)
Pool3D(2,2)	0.59 (±0.00)	**0.05** (±0.00)	0.88 (±0.00)	68.26 (±7.43)
Pool3D(3,1)	0.39 (±0.00)	0.02 (±0.00)	0.99 (±0.00)	117.52 (±9.36)
Pool3D(3,2)	0.59 (±0.00)	0.03 (±0.00)	0.66 (±0.00)	29.52 (±6.92)
Pool3D(4,1)	0.62 (±0.00)	0.03 (±0.00)	0.62 (±0.00)	39.30 (±0.49)

Table 19: MD hyperparameter searches for the MRI+ nnU-net (95% threshold). Dimensionality reduction techniques include PCA(n_p) with n_p components, UMAP(n_u) with n_u components, and n_d-dimensional average pooling with kernel size k and stride s, Pooln_dD(k,s). Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether higher or lower is better. Bold highlights the best performance per technique.

Experiment	**AUROC** ↑	**AUPRC** ↑	**FPR90** ↓	**Seconds** ↓
PCA(2)	0.88 (±0.00)	**1.00** (±0.00)	0.67 (±0.00)	1.04 (±0.04)
PCA(4)	0.65 (±0.00)	**1.00** (±0.00)	1.00 (±0.00)	1.17 (±0.06)
PCA(8)	**0.96** (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	1.12 (±0.04)
PCA(16)	0.91 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	1.16 (±0.03)
PCA(32)	0.89 (±0.00)	**1.00** (±0.00)	0.33 (±0.00)	1.22 (±0.03)
PCA(64)	0.81 (±0.01)	**1.00** (±0.00)	1.00 (±0.00)	1.28 (±0.08)
PCA(128)	0.85 (±0.01)	**1.00** (±0.00)	0.67 (±0.21)	1.66 (±0.05)
PCA(256)	0.84 (±0.01)	**1.00** (±0.00)	0.93 (±0.13)	2.27 (±0.08)
UMAP(2)	0.60 (±0.07)	0.99 (±0.00)	1.00 (±0.00)	18.34 (±0.71)
UMAP(4)	0.68 (±0.08)	**1.00** (±0.00)	0.87 (±0.27)	18.69 (±0.47)
UMAP(8)	0.77 (±0.07)	**1.00** (±0.00)	0.73 (±0.33)	19.27 (±0.94)
UMAP(16)	**0.82** (±0.08)	**1.00** (±0.00)	**0.67** (±0.03)	19.07 (±0.96)
UMAP(32)	0.66 (±0.09)	**1.00** (±0.00)	0.93 (±0.13)	18.74 (±0.63)
UMAP(64)	0.55 (±0.11)	**1.00** (±0.00)	0.93 (±0.13)	18.45 (±0.32)
UMAP(128)	0.64 (±0.03)	**1.00** (±0.00)	1.00 (±0.00)	19.20 (±0.85)
UMAP(256)	0.67 (±0.11)	**1.00** (±0.00)	1.00 (±0.00)	19.85 (±0.99)
Pool2D(2,1)	**0.85** (±0.00)	**1.00** (±0.00)	0.67 (±0.00)	1,579.73 (±52.41)
Pool2D(2,2)	0.49 (±0.00)	0.99 (±0.00)	1.00 (±0.00)	32.25 (±0.16)
Pool2D(3,1)	0.39 (±0.00)	0.99 (±0.00)	1.00 (±0.00)	372.05 (±7.89)
Pool2D(3,2)	0.20 (±0.00)	0.98 (±0.00)	1.00 (±0.00)	47.42 (±11.16)
Pool2D(4,1)	0.72 (±0.00)	**1.00** (±0.00)	**0.33** (±0.00)	103.14 (±20.14)
Pool3D(2,1)	0.36 (±0.00)	0.99 (±0.00)	1.00 (±0.00)	1,052.74 (±52.81)
Pool3D(2,2)	0.81 (±0.00)	**1.00** (±0.00)	0.67 (±0.00)	35.73 (±8.67)
Pool3D(3,1)	0.27 (±0.00)	0.99 (±0.00)	1.00 (±0.00)	145.59 (±13.95)
Pool3D(3,2)	0.84 (±0.00)	**1.00** (±0.00)	0.67 (±0.00)	24.64 (±11.49)
Pool3D(4,1)	0.66 (±0.00)	**1.00** (±0.00)	1.00 (±0.00)	36.02 (±11.78)

Table 20: MD hyperparameter searches for the CT nnU-net (95% threshold). Dimensionality reduction techniques include PCA(n_p) with n_p components, UMAP(n_u) with n_u components, and n_d-dimensional average pooling with kernel size k and stride s, Pooln_dD(k,s). Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether higher or lower is better. Bold highlights the best performance per technique.

Experiment	**AUROC ↑**	**AUPRC ↑**	**FPR90 ↓**	**Seconds ↓**
PCA(2)	0.45 (±0.00)	0.10 (±0.00)	0.95 (±0.00)	6.45 (±0.12)
PCA(4)	0.51 (±0.00)	0.16 (±0.00)	0.91 (±0.00)	6.64 (±0.11)
PCA(8)	0.55 (±0.00)	0.13 (±0.00)	**0.85** (±0.00)	7.22 (±0.07)
PCA(16)	0.51 (±0.00)	0.12 (±0.00)	0.96 (±0.00)	7.79 (±0.15)
PCA(32)	**0.56** (±0.00)	**0.17** (±0.00)	0.92 (±0.00)	8.17 (±0.23)
PCA(64)	**0.56** (±0.00)	0.16 (±0.00)	0.93 (±0.00)	8.57 (±0.24)
PCA(128)	**0.56** (±0.00)	0.15 (±0.00)	0.97 (±0.00)	9.53 (±0.36)
PCA(256)	0.54 (±0.00)	0.13 (±0.00)	0.98 (±0.00)	40.72 (±2.68)
UMAP(2)	0.58 (±0.07)	0.17 (±0.05)	0.86 (±0.11)	204.88 (±1.76)
UMAP(4)	0.52 (±0.02)	0.17 (±0.05)	0.93 (±0.02)	208.06 (±6.61)
UMAP(8)	0.57 (±0.03)	0.16 (±0.05)	0.87 (±0.07)	216.04 (±8.45)
UMAP(16)	0.63 (±0.02)	0.27 (±0.01)	0.77 (±0.02)	221.97 (±14.13)
UMAP(32)	0.67 (±0.01)	**0.29** (±0.03)	0.81 (±0.08)	268.08 (±25.20)
UMAP(64)	0.67 (±0.02)	0.26 (±0.02)	0.78 (±0.04)	305.58 (±46.07)
UMAP(128)	**0.68** (±0.03)	0.22 (±0.01)	0.74 (±0.09)	288.42 (±25.99)
UMAP(256)	0.62 (±0.03)	0.17 (±0.04)	**0.72** (±0.10)	351.84 (±18.14)
Pool2D(2,1)	0.46 (± 0.00)	0.12 (±0.00)	0.92 (±0.00)	1,886.94 (±51.54)
Pool2D(2,2)	**0.59** (±0.00)	**0.13** (±0.00)	0.84 (±0.00)	163.84 (±19.54)
Pool2D(3,1)	0.47 (±0.00)	0.11 (±0.00)	0.94 (±0.00)	654.53 (±34.88)
Pool2D(3,2)	0.46 (±0.00)	0.10 (±0.00)	0.92 (±0.00)	77.30 (±39.40)
Pool2D(4,1)	0.57 (±0.00)	**0.13** (±0.00)	0.73 (±0.00)	198.10 (±50.88)
Pool3D(2,1)	0.54 (±0.00)	0.12 (±0.00)	0.88 (±0.00)	1214.52 (±77.92)
Pool3D(2,2)	0.55 (±0.00)	0.11 (±0.00)	0.75 (±0.00)	106.38 (±27.31)
Pool3D(3,1)	0.49 (±0.00)	**0.13** (±0.00)	0.91 (±0.00)	298.91 (±15.05)
Pool3D(3,2)	0.50 (±0.00)	0.10 (±0.00)	0.90 (±0.00)	93.39 (±15.41)
Pool3D(4,1)	0.57 (±0.00)	**0.13** (±0.00)	**0.69** (±0.00)	125.90 (±13.68)

B.2 K-Nearest Neighbors

Table 21: KNN hyperparameter searches for MRI UNETR (95% threshold). Dimensionality reduction techniques include PCA(n_p) with n_p components, UMAP(n_u) with n_u components, and n_d-dimensional average pooling with kernel size k and stride s, Pooln_dD(k,s). Only the best-performing configuration of k is reported, with the logs containing all results available at `https://github.com/mckellwoodland/dimen_reduce_mahal/tree/main/logs`. Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether higher or lower is better. Bold highlights the best performance per technique.

Experiment	**K**	**AUROC ↑**	**AUPRC ↑**	**FPR90 ↓**	**Seconds ↓**
PCA(2)	256	**0.90** (±0.00)	**0.92** (±0.00)	0.31 (±0.00)	0.95 (±0.05)
PCA(4)	256	0.84 (±0.00)	0.88 (±0.00)	0.46 (±0.00)	0.93 (±0.04)
PCA(8)	256	0.84 (±0.00)	0.86 (±0.00)	0.43 (±0.04)	0.91 (±0.02)
PCA(16)	256	0.86 (±0.00)	0.88 (±0.00)	0.31 (±0.00)	1.04 (±0.08)
PCA(32)	256	0.86 (±0.00)	0.88 (±0.00)	0.31 (±0.00)	1.24 (±0.19)
PCA(64)	256	0.87 (±0.00)	0.88 (±0.00)	**0.23** (±0.00)	1.25 (±0.04)
PCA(128)	256	0.87 (±0.00)	0.88 (±0.00)	**0.23** (±0.00)	1.86 (±0.03)
PCA(256)	8	0.88 (±0.00)	0.89 (±0.00)	0.26 (±0.04)	2.64 (±0.08)
UMAP(2)	256	0.75 (±0.03)	0.79 (±0.07)	0.82 (±0.08)	5.67 (±0.13)
UMAP(4)	256	0.82 (±0.04)	0.84 (±0.04)	0.59 (±0.15)	6.41 (±0.20)
UMAP(8)	256	0.82 (±0.03)	0.82 (±0.06)	**0.46** (±0.10)	6.15 (±0.16)
UMAP(16)	256	0.79 (±0.03)	0.82 (±0.04)	0.57 (±0.08)	5.72 (±0.15)
UMAP(32)	256	**0.83** (±0.05)	**0.85** (±0.04)	0.51 (±0.08)	6.70 (±0.15)
UMAP(64)	256	0.80 (±0.05)	0.84 (±0.05)	0.65 (±0.13)	6.60 (±0.21)
UMAP(128)	256	0.77 (±0.07)	0.81 (±0.06)	0.69 (±0.17)	6.80 (±0.13)
UMAP(256)	256	0.80 (±0.04)	0.84 (±0.03)	0.74 (±0.17)	7.21 (±0.12)
Pool2D(2,1)	256	**0.98** (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	4.90 (±0.06)
Pool2D(2,2)	256	**0.98** (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	0.74 (±0.04)
Pool2D(3,1)	256	0.96 (±0.00)	**1.00** (±0.00)	0.33 (±0.00)	2.75 (±0.03)
Pool2D(3,2)	256	0.94 (±0.00)	**1.00** (±0.00)	0.33 (±0.00)	9.54 (±8.16)
Pool2D(4,1)	256	0.91 (±0.00)	**1.00** (±0.00)	0.33 (±0.00)	1.09 (±0.01)
Pool3D(2,1)	256	0.96 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	4.27 (±0.08)
Pool3D(2,2)	256	0.85 (±0.00)	**1.00** (±0.00)	1.00 (±0.00)	0.36 (±0.01)
Pool3D(3,1)	256	0.94 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	1.76 (±0.08)
Pool3D(3,2)	256	0.92 (±0.00)	**1.00** (±0.00)	0.33 (±0.00)	0.36 (±0.00)
Pool3D(4,1)	256	0.89 (±0.00)	**1.00** (±0.00)	0.33 (±0.00)	0.64 (±0.02)

Table 22: KNN hyperparameter searches for MRI+ UNETR (80% threshold). Dimensionality reduction techniques include PCA(n_p) with n_p components, UMAP(n_u) with n_u components, and n_d-dimensional average pooling with kernel size k and stride s, Pooln_dD(k,s). Only the best-performing configuration of k is reported, with the logs containing all results available at `https://github.com/mckellwoodland/dimen_reduce_mahal/tree/main/logs`. Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether higher or lower is better. Bold highlights the best performance per technique.

Experiment	**K**	**AUROC ↑**	**AUPRC ↑**	**FPR90 ↓**	**Seconds ↓**
PCA(2)	128	0.62 (±0.00)	0.03 (±0.00)	0.49 (±0.00)	1.43 (±0.10)
PCA(4)	128	0.63 (±0.00)	0.03 (±0.00)	0.54 (±0.00)	1.35 (±0.03)
PCA(8)	128	0.78 (±0.00)	0.06 (±0.00)	0.43 (±0.00)	1.42 (±0.04)
PCA(16)	64	0.82 (±0.01)	0.09 (±0.01)	0.42 (±0.01)	1.44 (±0.03)
PCA(32)	64	**0.84** (±0.01)	**0.17** (±0.02)	**0.40** (±0.01)	1.57 (±0.06)
PCA(64)	128	0.83 (±0.01)	0.11 (±0.01)	0.42 (±0.01)	1.67 (±0.10)
PCA(128)	128	0.83 (±0.00)	0.13 (±0.01)	0.42 (±0.01)	2.66 (±0.11)
PCA(256)	64	0.83 (±0.00)	0.14 (±0.01)	0.42 (±0.01)	4.26 (±1.19)
UMAP(2)	128	0.69 (±0.09)	0.07 (±0.06)	0.46 (±0.08)	17.26 (±0.61)
UMAP(4)	64	**0.78** (±0.05)	0.09 (±0.06)	0.42 (±0.07)	15.70 (±0.45)
UMAP(8)	128	0.73 (±0.09)	0.06 (±0.03)	0.46 (±0.06)	17.80 (±0.26)
UMAP(16)	128	0.73 (±0.05)	0.08 (±0.06)	0.46 (±0.03)	17.26 (±0.77)
UMAP(32)	128	0.73 (±0.07)	0.06 (±0.04)	0.46 (±0.05)	17.23 (±0.25)
UMAP(64)	256	0.74 (±0.07)	0.05 (±0.01)	0.42 (±0.06)	17.43 (±0.88)
UMAP(128)	256	0.74 (±0.05)	0.05 (±0.02)	0.44 (±0.03)	18.50 (±0.86)
UMAP(256)	64	0.75 (±0.06)	**0.12** (±0.16)	**0.39** (±0.03)	17.27 (±0.50)
Pool2D(2,1)	128	0.86 (±0.00)	0.19 (±0.00)	0.40 (±0.00)	6.28 (±0.05)
Pool2D(2,2)	256	0.85 (±0.00)	0.27 (±0.00)	0.43 (±0.00)	13.53 (±7.27)
Pool2D(3,1)	64	0.86 (±0.00)	0.15 (±0.00)	0.31 (±0.00)	2.81 (±0.06)
Pool2D(3,2)	256	0.76 (±0.00)	0.07 (±0.00)	0.58 (±0.00)	1.34 (±0.06)
Pool2D(4,1)	32	0.85 (±0.00)	0.27 (±0.00)	0.40 (±0.00)	9.09 (±5.19)
Pool3D(2,1)	256	**0.87** (±0.00)	0.30 (±0.00)	0.38 (±0.00)	23.34 (±2.74)
Pool3D(2,2)	256	**0.87** (±0.00)	**0.33** (±0.00)	0.43 (±0.00)	11.21 (±7.66)
Pool3D(3,1)	64	**0.87** (±0.00)	0.19 (±0.00)	**0.30** (±0.00)	2.30 (±0.10)
Pool3D(3,2)	64	0.78 (±0.00)	0.06 (±0.00)	0.41 (±0.00)	11.14 (±6.26)
Pool3D(4,1)	8	0.84 (±0.00)	0.26 (±0.00)	0.41 (±0.00)	1.24 (±0.07)

Table 23: KNN hyperparameter searches for MRI+ nnU-net (95% threshold). Dimensionality reduction techniques include PCA(n_p) with n_p components, UMAP(n_u) with n_u components, and n_d-dimensional average pooling with kernel size k and stride s, Pooln_dD(k,s). Only the best-performing configuration of k is reported, with the logs containing all results available at `https://github.com/mckellwoodland/dimen_reduce_mahal/tree/main/logs`. Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether higher or lower is better. Bold highlights the best performance per technique.

Experiment	**K**	**AUROC** ↑	**AUPRC** ↑	**FPR90** ↓	**Seconds** ↓
PCA(2)	256	0.94 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	0.99 (±0.02)
PCA(4)	256	0.75 (±0.00)	**1.00** (±0.00)	1.00 (±0.00)	1.04 (±0.02)
PCA(8)	256	0.97 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	1.16 (±0.07)
PCA(16)	256	**0.98** (±0.00)	**1.00** (±0.00)	0.33 (±0.00)	1.16 (±0.08)
PCA(32)	256	0.95 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	1.25 (±0.04)
PCA(64)	256	0.95 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	1.24 (±0.07)
PCA(128)	256	0.96 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	1.71 (±0.07)
PCA(256)	256	0.95 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	2.89 (±0.41)
UMAP(2)	2	**0.96** (±0.04)	**1.00** (±0.00)	**0.13** (±0.16)	14.91 (±2.08)
UMAP(4)	2	0.94 (±0.02)	**1.00** (±0.00)	0.33 (±0.00)	15.15 (±2.30)
UMAP(8)	2	0.94 (±0.03)	**1.00** (±0.00)	0.40 (±0.33)	15.22 (±1.57)
UMAP(16)	2	0.94 (±0.02)	**1.00** (±0.00)	0.26 (±0.13)	14.89 (±2.38)
UMAP(32)	2	0.89 (±0.05)	**1.00** (±0.00)	0.40 (±0.33)	14.64 (±1.92)
UMAP(64)	2	0.92 (±0.02)	**1.00** (±0.00)	0.26 (±0.13)	14.82 (±1.92)
UMAP(128)	2	0.92 (±0.02)	**1.00** (±0.00)	0.33 (±0.00)	15.08 (±1.74)
UMAP(256)	2	0.91 (±0.04)	**1.00** (±0.00)	0.40 (±0.14)	15.75 (±1.83)
Pool2D(2,1)	256	**0.98** (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	4.90 (±0.06)
Pool2D(2,2)	256	**0.98** (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	0.74 (±0.04)
Pool2D(3,1)	256	0.96 (±0.00)	**1.00** (±0.00)	0.33 (±0.00)	2.75 (±0.03)
Pool2D(3,2)	256	0.94 (±0.00)	**1.00** (±0.00)	0.33 (±0.00)	9.54 (±8.16)
Pool2D(4,1)	256	0.91 (±0.00)	**1.00** (±0.00)	0.33 (±0.00)	1.09 (±0.01)
Pool3D(2,1)	256	0.96 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	4.27 (±0.08)
Pool3D(2,2)	256	0.85 (±0.00)	**1.00** (±0.00)	1.00 (±0.00)	0.36 (±0.01)
Pool3D(3,1)	256	0.94 (±0.00)	**1.00** (±0.00)	**0.00** (±0.00)	1.76 (±0.08)
Pool3D(3,2)	256	0.92 (±0.00)	**1.00** (±0.00)	0.33 (±0.00)	0.36 (±0.00)
Pool3D(4,1)	256	0.89 (±0.00)	**1.00** (±0.00)	0.33 (±0.00)	0.64 (±0.02)

Table 24: KNN hyperparameter searches for CT nnU-net (95% threshold). Dimensionality reduction techniques include PCA(n_p) with n_p components, UMAP(n_u) with n_u components, and n_d-dimensional average pooling with kernel size k and stride s, Pooln_dD(k,s). Only the best-performing configuration of k is reported, with the logs containing all results available at `https://github.com/mckellwoodland/dimen_reduce_mahal/tree/main/logs`. Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether higher or lower is better. Bold highlights the best performance per technique.

Experiment	**K**	**AUROC** ↑	**AUPRC** ↑	**FPR90** ↓	**Seconds** ↓
PCA(2)	8	0.42 (±0.00)	0.09 (±0.00)	0.93 (±0.00)	4.95 (±0.12)
PCA(4)	256	0.47 (±0.00)	0.11 (±0.00)	**0.91** (±0.00)	5.91 (±0.84)
PCA(8)	4	**0.55** (±0.00)	**0.15** (±0.00)	0.97 (±0.00)	4.94 (±0.04)
PCA(16)	4	0.52 (±0.00)	0.12 (±0.00)	0.96 (±0.00)	5.00 (±0.04)
PCA(32)	4	0.52 (±0.00)	0.13 (±0.00)	0.97 (±0.00)	5.20 (±0.07)
PCA(64)	8	0.53 (±0.00)	0.13 (±0.00)	0.96 (±0.00)	5.60 (±0.10)
PCA(128)	4	0.53 (±0.00)	0.13 (±0.00)	0.96 (±0.00)	5.98 (±0.04)
PCA(256)	4	0.53 (±0.00)	0.13 (±0.00)	0.96 (±0.00)	24.75 (±14.29)
UMAP(2)	256	0.64 (±0.02)	**0.24** (±0.04)	0.88 (±0.03)	200.58 (±1.93)
UMAP(4)	256	**0.65** (±0.01)	**0.24** (±0.05)	0.88 (±0.02)	199.93 (±1.96)
UMAP(8)	256	**0.65** (±0.01)	0.22 (±0.04)	0.87 (±0.02)	198.33 (±2.34)
UMAP(16)	256	**0.65** (±0.00)	0.23 (±0.05)	**0.86** (±0.01)	204.01 (±1.27)
UMAP(32)	256	0.63 (±0.01)	0.19 (±0.04)	0.88 (±0.02)	195.59 (±0.55)
UMAP(64)	256	**0.65** (±0.01)	0.22 (±0.03)	0.88 (±0.01)	197.23 (±0.68)
UMAP(128)	256	0.64 (±0.01)	0.18 (±0.01)	**0.86** (±0.03)	199.48 (±1.40)
UMAP(256)	256	0.64 (±0.00)	0.19 (±0.02)	0.87 (±0.03)	200.73 (±0.65)
Pool2D(2,1)	4	0.53 (±0.00)	**0.15** (±0.00)	0.95 (±0.00)	97.91 (±9.21)
Pool2D(2,2)	32	0.52 (±0.00)	**0.15** (±0.00)	0.95 (±0.00)	33.17 (±6.84)
Pool2D(3,1)	2	0.52 (±0.00)	0.14 (±0.00)	**0.93** (±0.00)	104.41 (±14.69)
Pool2D(3,2)	4	0.51 (±0.00)	0.12 (±0.00)	0.95 (±0.00)	84.06 (±16.41)
Pool2D(4,1)	4	0.52 (±0.00)	0.11 (±0.00)	0.95 (±0.00)	42.95 (±16.59)
Pool3D(2,1)	4	**0.54** (±0.00)	0.14 (±0.00)	0.96 (±0.00)	98.68 (±39.50)
Pool3D(2,2)	4	**0.54** (±0.00)	0.14 (±0.00)	0.96 (±0.00)	81.96 (±26.79)
Pool3D(3,1)	4	0.53 (±0.00)	0.14 (±0.00)	0.94 (±0.00)	81.41 (±28.32)
Pool3D(3,2)	4	0.53 (±0.00)	0.13 (±0.00)	0.95 (±0.00)	75.00 (±14.21)
Pool3D(4,1)	4	0.53 (±0.00)	0.13 (±0.00)	**0.93** (±0.00)	84.91 (±11.64)

B.3 Temperature Scaling and Energy Scoring

Table 25: Temperature scaling (TS) and energy scoring (Energy) hyperparameter searches for the UNETRs. Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether higher or lower is better. Bold denotes the best performance per method and model.

Model	Method	T	AUROC ↑	AUPRC ↑	FPR90 ↓	Seconds ↓
MRI UNETR	TS	2	**0.90** (±0.00)	**0.93** (±0.00)	**0.15** (±0.00)	9.25 (±0.11)
		3	0.72 (±0.00)	0.77 (±0.00)	0.38 (±0.00)	9.22 (±0.04)
		4	0.67 (±0.00)	0.71 (±0.00)	0.46 (±0.00)	9.21 (±0.10)
		5	0.61 (±0.00)	0.64 (±0.00)	0.54 (±0.00)	9.08 (±0.06)
		10	0.57 (±0.00)	0.58 (±0.00)	0.62 (±0.00)	9.17 (±0.06)
		100	0.55 (±0.00)	0.57 (±0.00)	0.69 (±0.00)	9.22 (±0.10)
		1000	0.55 (±0.00)	0.57 (±0.00)	0.69 (±0.00)	9.20 (±0.07)
	Energy	1	**0.55** (±0.00)	**0.57** (±0.00)	**0.69** (±0.00)	7.32 (±0.28)
		2	**0.55** (±0.00)	**0.57** (±0.00)	**0.69** (±0.00)	7.25 (±0.17)
		3	**0.55** (±0.00)	**0.57** (±0.00)	**0.69** (±0.00)	**7.08** (±0.14)
		4	**0.55** (±0.00)	**0.57** (±0.00)	**0.69** (±0.00)	7.21 (±0.12)
		5	**0.55** (±0.00)	**0.57** (±0.00)	**0.69** (±0.00)	7.38 (±0.22)
		10	0.54 (±0.00)	**0.57** (±0.00)	**0.69** (±0.00)	7.11 (±0.09)
		100	0.54 (±0.00)	**0.57** (±0.00)	**0.69** (±0.00)	7.28 (±0.11)
		1000	0.54 (±0.00)	**0.57** (±0.00)	**0.69** (±0.00)	7.19 (±0.14)
MRI+ UNETR	TS	2	**0.47** (±0.00)	**0.05** (±0.00)	**0.78** (±0.00)	60.57 (±0.31)
		3	0.44 (±0.00)	0.04 (±0.00)	0.86 (±0.00)	60.87 (±0.27)
		4	0.42 (±0.00)	0.04 (±0.00)	0.88 (±0.00)	60.55 (±0.27)
		5	0.42 (±0.00)	0.03 (±0.00)	0.90 (±0.00)	60.56 (±0.16)
		10	0.42 (±0.00)	0.03 (±0.00)	0.89 (±0.00)	60.15 (±0.12)
		100	0.44 (±0.00)	0.04 (±0.00)	0.88 (±0.00)	60.79 (±0.75)
		1000	0.45 (±0.00)	0.04 (±0.00)	0.88 (±0.00)	60.67 (±0.18)
	Energy	1	0.44 (±0.00)	0.03 (±0.00)	0.89 (±0.00)	35.39 (±1.46)
		2	0.43 (±0.00)	0.03 (±0.00)	0.90 (±0.00)	**34.82** (±0.38)
		3	0.44 (±0.00)	0.03 (±0.00)	0.89 (±0.00)	34.94 (±0.23)
		4	0.46 (±0.00)	0.03 (±0.00)	0.88 (±0.00)	35.14 (±0.58)
		5	0.47 (±0.00)	0.04 (±0.00)	0.87 (±0.00)	35.61 (±0.40)
		10	0.51 (±0.00)	0.05 (±0.00)	0.85 (±0.00)	36.23 (±0.11)
		100	0.54 (±0.00)	**0.06** (±0.00)	0.79 (±0.00)	35.92 (±0.63)
		1000	**0.61** (±0.00)	0.03 (±0.00)	**0.71** (±0.00)	35.86 (±0.41)

Table 26: Temperature scaling (TS) and energy scoring (Energy) hyperparameter searches for the nnU-nets. NaN signifies that the calculation did not produce a number due to computational instabilities. Seconds is the amount of time it took to calculate the test distances. The results are averages (±SD) across 5 runs. Arrows denote whether higher or lower is better. Bold denotes the best performance per method and model.

Model	**Method**	**T**	**AUROC ↑**	**AUPRC ↑**	**FPR90 ↓**	**Seconds ↓**
MRI+ nnU-net	TS	2	0.45 (±0.00)	**0.99** (±0.00)	**1.00** (±0.00)	1,254.10 (±0.29)
		3	0.50 (±0.00)	**0.99** (±0.00)	**1.00** (±0.00)	1,252.72 (±0.52)
		4	0.53 (±0.00)	**0.99** (±0.00)	**1.00** (±0.00)	1,253.29 (±0.62)
		5	0.53 (±0.00)	**0.99** (±0.00)	**1.00** (±0.00)	1,252.59 (±0.76)
		10	**0.55** (±0.00)	**0.99** (±0.00)	**1.00** (±0.00)	1,252.78 (±0.73)
		100	0.50 (±0.00)	**0.99** (±0.00)	**1.00** (±0.00)	1,253.07 (±0.80)
		1000	0.50 (±0.00)	**0.99** (±0.00)	**1.00** (±0.00)	**1,204.30** (±0.48)
	Energy	1	0.58 (±0.00)	**1.00** (±0.00)	**1.00** (±0.00)	186.32 (±0.49)
		2	0.58 (±0.00)	**1.00** (±0.00)	**1.00** (±0.00)	**185.73** (±0.76)
		3	0.59 (±0.00)	**1.00** (±0.00)	**1.00** (±0.00)	186.50 (±0.49)
		4	0.59 (±0.00)	**1.00** (±0.00)	**1.00** (±0.00)	186.55 (±0.76)
		5	0.60 (±0.00)	**1.00** (±0.00)	**1.00** (±0.00)	186.62 (±0.19)
		10	**0.61** (±0.00)	**1.00** (±0.00)	**1.00** (±0.00)	186.88 (±0.47)
		100	NaN	NaN	NaN	NaN
		1000	NaN	NaN	NaN	NaN
CT nnU-net	TS	2	**0.68** (±0.00)	0.25 (±0.00)	0.66 (±0.00)	699.65 (±0.83)
		3	**0.68** (±0.00)	**0.26** (±0.00)	0.69 (±0.00)	699.07 (±0.37)
		4	**0.68** (±0.00)	0.25 (±0.00)	0.69 (±0.00)	699.38 (±0.60)
		5	0.67 (±0.00)	0.25 (±0.00)	0.69 (±0.00)	699.45 (±0.39)
		10	0.67 (±0.00)	0.24 (±0.00)	0.70 (±0.00)	699.68 (±0.42)
		100	0.67 (±0.00)	0.15 (±0.00)	**0.36** (±0.00)	700.37 (±0.41)
		1000	0.57 (±0.00)	0.10 (±0.00)	1.00 (±0.00)	**670.35** (±0.69)
	Energy	1	0.66 (±0.00)	**0.26** (±0.00)	**0.72** (±0.00)	105.88 (±0.90)
		2	**0.67** (±0.00)	**0.26** (±0.00)	0.75 (±0.00)	**105.66** (±0.60)
		3	**0.67** (±0.00)	0.25 (±0.00)	0.75 (±0.00)	106.26 (±0.52)
		4	0.66 (±0.00)	0.24 (±0.00)	0.75 (±0.00)	106.06 (±0.68)
		5	0.66 (±0.00)	0.23 (±0.00)	0.75 (±0.00)	106.15 (±0.61)
		10	0.66 (±0.00)	0.17 (±0.00)	0.76 (±0.00)	106.12 (±0.56)
		100	NaN	NaN	NaN	NaN
		1000	0.40 (±0.00)	0.08 (±0.00)	1.00 (±0.00)	105.69 (±0.62)

Appendix C. Additional Figures

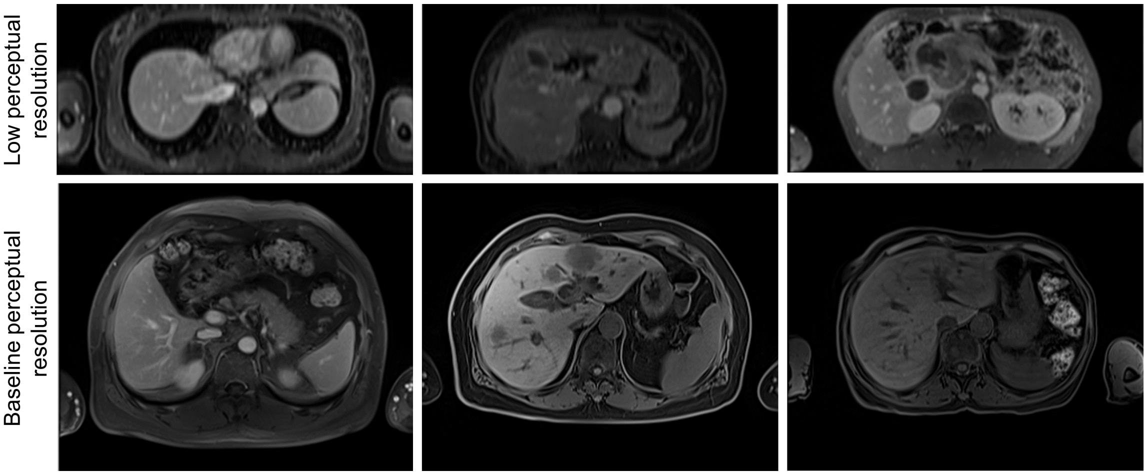


Figure 5: Sample images from the AMOS dataset (Ji, 2022) that represent the baseline and low perceptual resolution images that were clustered separately by the dimensionality reduction techniques.

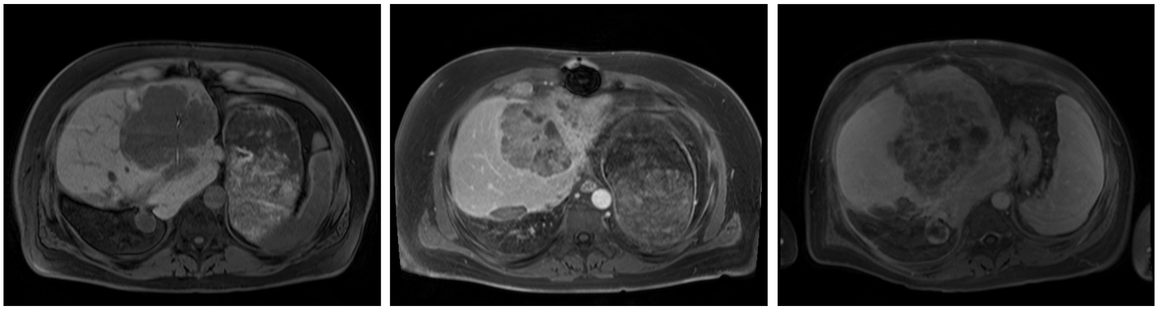


Figure 6: Sample slices from different scans from the same patient with a large tumor.