Anomaly detection and segmentation methods using Variational Autoencoders

Major Research Project

MSc Medical Imaging

Sara Guillén Fernández-Micheltorena

Examination committee:

Supervisor: dr. ir. Koen Vincken

Assistant professor, UMC Utrecht

Examiner: dr. Matteo Maspero

Assistant professor, UMC Utrecht

Utrecht, 21st February 2023

Anomaly Detection and Localization Methods using Variational Autoencoders

Sara Guillén Fernández-Micheltorena
Imaging and Oncology Division
University Medical Center in Utrecht (UMCU)
s.guillenfernandezmicheltorena@students.uu.nl

***Abstract*—Variational Autoencoders (VAEs) have emerged as a promising technique for unsupervised anomaly detection and segmentation in brain MRI. The principle behind this deep generative modeling is learning a model of healthy brain anatomy by encoding this information into a latent representation and reconstructing it. Anomalies can be detected and localized by discrepancies between the reconstructed data and the original input. This technique relieves the need for pixel-level segmented data and provides the possibility to detect arbitrary anomalies. Challenges have been reported in recent publications regarding the sensitivity towards bright lesions and the limitations of commonly used anomaly scoring methods based on reconstruction error and residual images. These issues are addressed in this work by changing the background of brain MR scans, and introducing anomaly scoring methods based on outlier detection, and activation maps. The evaluation of these methods is performed using four synthetic datasets containing toy anomalies varying in size, intensity, and number. The results show the possibility of detecting dark lesions, the potential of outlier detection in latent space for anomaly detection, and the extraction of activation maps for anomaly segmentation.**

***Index Terms*—Anomaly detection, anomaly segmentation, Variational Autoencoders, Brain MRI, anomaly scoring**

1. Introduction

Brain MRI is of utmost relevance when diagnosing and treating neurological disorders. Radiologists are medical professionals who read and interpret MR studies, a complex process susceptible to errors. These diagnostic errors range from 3% to 5% [1]. Rapid advances in Artificial Intelligence have seen the emergence of supervised deep learning algorithms to assist radiologists in improving the accuracy and speed of medical diagnoses [2]. Despite the outstanding performances achieved by these models, their training requires an extensive amount of labeled data which is scarce and time-consuming to produce. Moreover, the generalization of these methods is limited to the available anomalies present in training data. Recently, unsupervised anomaly detection and segmentation have attracted attention in the field of medical image analysis with two main advantages: they do not require pixel-level annotations and can detect arbitrary pathologies.

A common technique to tackle unsupervised anomaly detection and segmentation is through deep generative modeling using Variational Autoencoders (VAEs) [3]. The principle behind this approach relies on learning a model of healthy brain anatomy by deep representation learning which can reconstruct the original data. After training, anomalies and anomalous regions can be detected by comparing the original data and their reconstructed normal counterpart.

In the field of medical imaging, current literature employs VAEs [4]–[9]. The study by Baur et al. [10] reviews unsupervised anomaly segmentation methods based on autoencoders. However, the results from Medical Out-of Distribution (MOOD) challenge in [11] and the comparison to a simple thresholding method in [12] show their poor performance which might be impractical for clinical practice. In fact, these methods are sensitive to bright lesions [12]. The anomaly scoring method often used is based on the reconstruction image generated by the decoder, both the reconstruction error for anomaly detection and the residual image for anomaly segmentation. However, as discussed in [13], intensity and texture variations in anomalies can interfere with the correct reconstruction making residual images unreliable. Recently, new scoring methods have been investigated such as in [14] where authors exploit the deep representation of the latent space to detect outliers, or in [15] where activation maps are used to segment anomalous regions.

The principal contributions of this work that address these issues can be summarized in the following:

- We propose an extra pre-processing step in brain MR images to reduce the sensitivity to bright lesions by changing the background.
- We investigate outlier detection in latent space as an alternative to the reconstruction error technique for anomaly detection.
- We compare the computation of residual images to the extraction of activation maps as anomaly segmentation methods.
- We propose an evaluation of all the experiments using toy anomalies with varying sizes, intensities, shapes, and number.

The methods implemented in this work are explored in Section 2. The datasets and corresponding experiments are detailed in Section 3. The results and discussion are explained in Sections 4 and 5, respectively. The final conclusion corresponds to Section 6.

2. Methods

2.1. Simple VAE

A VAE is a generative model that can be used to learn a latent distribution of a dataset. It consists of an encoder and a decoder. The encoder network maps the input data into a lower-dimensional representation (also known as latent space). This representation is given by μ and σ [16]. Then, a point from the latent space is sampled, and the decoder network generates an approximation of the original data [16]. Training a VAE consists in minimizing a two-term loss function, which is equivalent to maximizing the evidence lower-bound (ELBO) [17]. This can be written as:

$$L_{VAE} = L_R(x, \hat{x}) + \beta L_{KL}(q(z|x)||p(z))$$

where x is the input image, z its latent representation and $\hat{x}$ the reconstructed counterpart. L_R is the reconstruction error term between the input image and its reconstruction. L_{KL} is the Kullback- Leibler (KL) divergence (weighted by β) which measures the difference between two probability distributions: the approximate posterior $q(z|x)$ and the prior $p(z)$. The prior distribution is often set as the standard normal distribution with a mean of zero and a standard deviation of one. The first term encourages the VAE to reconstruct the input data accurately, while the second term encourages the latent representation to be close to the prior (standard normal) distribution.

The KL loss is calculated as the relative entropy between the latent reconstruction that comes from the distribution given by μ and σ and the standard normal distribution:

$$L_{KL} = -\frac{1}{2}\sum_{j=1}^{J}(1 + log(\sigma_j^2) - \mu_j^2 - \sigma_j^2)$$

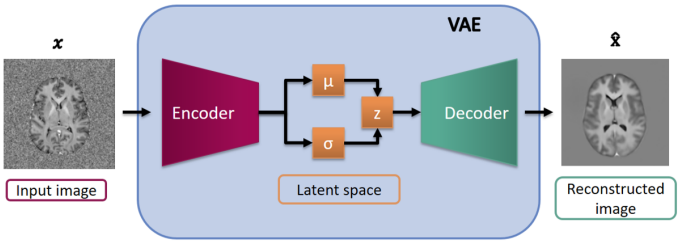

Fig. 1: Diagram of the components of a VAE. The input image x is compressed by the encoder into the latent space. The sampled latent vector is decoded resulting in the reconstruction image.

In this work, a simple VAE is implemented for the experiments using the Fashion MNIST dataset. The encoder of the simple VAE consists of 4 convolutional layers with ReLU activation, each followed by a batch normalization layer. The bottleneck of the network has two separate fully connected layers leading to a dense latent space of 16 dimensions. For the image generation part, the decoder is symmetrical to the encoder. The reconstruction error is the binary cross-entropy (BCE). The optimizer is Adam with a learning rate of 1e-4 and weight decay of 1e-5. The chosen batch size is 64.

2.2. Constrained VAE

For the application of non-weighted activation maps (AMs) for anomaly segmentation, an extra regularization term in the loss function of a VAE is included [15]. This term aims to regularize the attention distribution in order to focus on the whole image patterns homogeneously by maximizing the entropy of the said distribution. The loss function of the constrained VAE can then be written as:

$$L_{ConsVAE} = L_{VAE} + L_H(H(a))$$

Where the left-hand term is the standard VAE loss function and L_H is the entropy regularizer term which calculates the entropy of the activation map a.

The constrained VAE architecture used is based on [15]. The encoder has the convolutional layers of ResNet-18 [18] followed by a dense latent space. Lastly, the residual decoder is symmetrical to the encoder. The reconstruction loss is computed using the mean square error (MSE).

This architecture is used to train two models using the MOOD challenge dataset: one with 2D slices and one with 2D patches. The latent space dimension is set to 128 for the training with slices and 64 for the training of patches. For the slices, the optimizer is Adam with a learning rate of 1e-4 and weight decay of 1e-5. For patches, the optimizer is Adam with a learning rate of 5e-5 and weight decay of 1e-5. The chosen batch size for both cases is 64.

2.3. Anomaly scoring

The anomaly scoring methods in anomaly detection determine whether the input sample is normal or anomalous, giving a single score from 0 (normal) to 1 (anomalous) to the sample. In the case of anomaly segmentation, each pixel of the sample receives a score from 0 to 1, which results in an anomaly map. These anomaly maps are thresholded creating an anomaly segmentation mask.

The anomaly scoring methods for anomaly detection investigated in this work are the following:

- Reconstruction error: It consists of computing the MSE between the input sample (x) and the reconstructed sample ($\hat{x}$) generated by the decoder. It is assumed that the decoder will better reconstruct a sample that falls into the distribution learned during training. Therefore, a low reconstruction error would correspond to a normal sample while a high reconstruction error would refer to an anomalous sample.

$$Score = MSE(x, \hat{x})$$

- Outlier detection in latent space: It consists of generating a decision function (F) in the learned latent space corresponding to the distribution of normal samples. Then, a new sample is encoded (z) and the decision function computes an outlier score (anomaly score). The outlier

detection algorithms are implemented using the Pyod package in Python [19] from which 16 outlier detection methods are chosen (see Appendix A) to be tested.

$$Score = F(z)$$

In the anomaly segmentation case, the methods explored are:

- Residual image: The residual image is computed by subtracting the input image and the reconstructed normal counterpart generated by the decoder. The resulting image corresponds to the anomaly present in the input image which is not generated by the VAE in the reconstructed one.

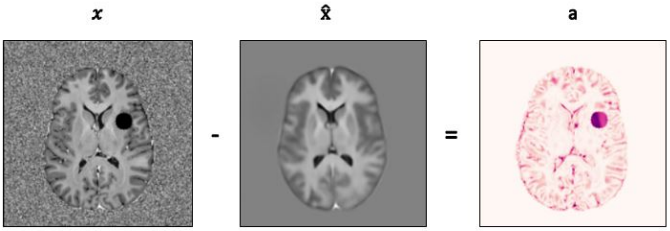

Fig. 2: Example of the computation of the residual image using the input (x) and reconstructed ($\hat{x}$) images.

- Activation map: These maps are obtained by extracting activation maps from layers of the encoder. Large activation values correspond to anomalies in the image [15].

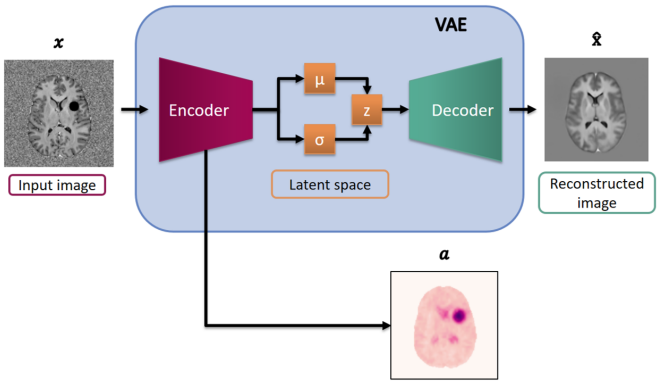

Fig. 3: Example of the extraction of the activation map from the layers of the encoder of a trained VAE.

These anomaly detection and segmentation scoring methods are applied to both slices and patches. In the case of patches, the process involves dividing the test image into patches, applying the techniques, and then recombining the patches to generate an anomaly map. In the case of anomaly detection methods, each patch receives a single score. When recombining the patches to generate the initial slice, a coarse anomaly map is generated that is dependent on the stride used for patch extraction. Conversely, in the case of anomaly segmentation methods, each pixel of the patch receives an anomaly score, leading to a final anomaly map with the original resolution of the image. This allows an extra comparison of the methods using the model trained in slices and the model trained in patches.

2.4. Evaluation metrics

Current methods tend to choose the area under the receiver operating characteristic (AUROC) as the evaluation metric for anomaly detection. However, when positive examples are scarce and of utmost importance for detection, AUROC scores become an inadequate measure of detection performance, as they show optimistic results [20]. In such cases, precision becomes more meaningful than the false positive rate.

This is why, in this study, the metric implemented to evaluate the experiments using the different anomaly scoring methods is the average precision score (AP), as it was done in [13]. AP is calculated as the mean of the precision scores obtained at each threshold (P_n) weighted by the increase of the recall from the previous threshold as the weight ($R_n - R_{n-1}$):

$$AP = \sum_n (R_n - R_{n-1}) P_n$$

The AP score is an approximation of the area under the precision-recall curve (AUPRC). This avoids the election of a threshold for the output, integrating all thresholds in its calculation [21]. Moreover, contrary to AUROC, this metric is robust regarding class imbalance [20].

The evaluation of the different anomaly detection methods for both Fashion MNIST and MOOD datasets is assessed at a 2D image level, considering both normal and anomalous images. The anomaly segmentation performance for the MOOD dataset is assessed at pixel level per single slice. Finally, the average values are obtained for each 3D scan accounting for all the slices. The results for every one of the anomaly scoring methods will display the AP of each anomaly type (size and/or intensity-wise) present in the 4 toy image datasets explained in Section 3.2.4, as well as the mean AP (mAP) for each entire dataset.

The results will show the performance of each method for each type of anomaly using the AP. The results will also display the mean average precision (mAP) for each testing dataset, which is the average of all the AP scores for the anomalies in that dataset. These datasets are described in more detail in Sections 3.1 (Fashion MNIST) and 3.2.4 (MOOD).

3. Data & Experiments

3.1. Fashion MNIST

Fashion MNIST is a dataset with images of Zalando's articles [22]. It contains 60,000 training and 10,000 testing 28x28 greyscale images. Each sample has an associated label that corresponds to one of the 10 different items: t-shirt, trousers, pullover, dress, coat, sandal, shirt, sneaker, bag, and ankle boot. This dataset is intended to replace the original MNIST dataset as a benchmark for validating machine learning algorithms, as MNIST is overused and can easily achieve

an accuracy of 99.7% in convolutional networks and 97% in classic machine learning algorithms [22].

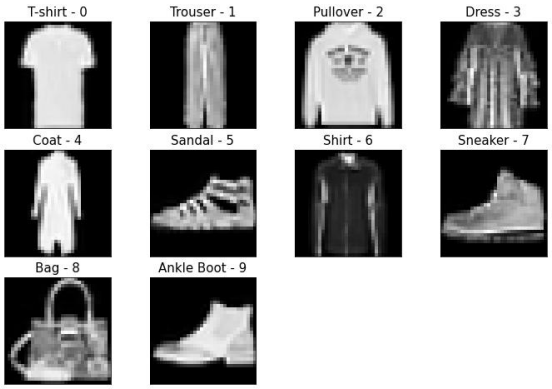

Fig. 4: Fashion MNIST image examples of each class and their corresponding labels.

3.1.1. Anomaly detection

For anomaly detection, the simple VAE model is trained with one item (t-shirt), which is considered "normal", and then tested with the other items, which are considered "anomalous". The images are rescaled to 32x32 because the input size for this VAE must be a multiple of 32, otherwise, information is lost. The 7,000 t-shirt images are divided into training, validation, and testing datasets in the following way: 70% or 4,900 images for training, 20% or 1,400 images for validation, and 10% or 700 images for testing. The other 63,000 images of the 9 anomalous classes are used for testing. The model is trained for 50 epochs.

The anomaly scoring methods for anomaly detection applied in this experiment are reconstruction error and outlier detection in latent space. The anomaly localization methods are not implemented in this experiment.

3.2. MOOD Challenge

The data available from the Medical Out of Distribution (MOOD) challenge [11] consists of two datasets, one with brain MRI images and one with abdominal CT scans. In this study, the brain MR dataset containing 800 256x256x256 images will be used. These images have been acquired with the same scan and show T1-weighted contrast. The images are already skull stripped. They also provide four toy images which consist of one original image with an inserted sphere of a certain intensity. Additionally, extra toy images are created containing spheres and random shapes varying in number, size, and intensity for the evaluation process.

3.2.1. Pre-processing

The MOOD images are pre-processed in 3 steps:

- Registration to the SRI24 atlas [23] to guarantee the same volume, size, and orientation of the data. The resulting size of the 3D images is 155x240x240.
- Normalization using a mean (μ) of 0 and standard deviation (σ) of 1 which results in 99.74% of the values to lay on a range -3 to 3. Then, the image values are divided by 3 and clipped so that the final values range from -1 to 1 with a mean close to 0.
- The background of the images is changed by the addition of noise with the same distribution as the brain region: $\mu = 0$ and $\sigma = 1/3$.

3.2.2. 2D axial slices

The 800 images are split into a training set of 560 (80%), a validation set of 160 (20%) and a test set of 80 (10%). Each image is divided into 2D axial slices. These slices are cropped to be 224x224, as the input images for this VAE should be multiples of 32, and the empty ones are removed. Cropping is chosen over resizing, as only background information is lost. 50,000 slices are used for training and 5,000 for validation. The test set is used to create anomalies for the evaluation.

The anomaly scoring methods for anomaly detection in 2D slices are the reconstruction error and outlier detection in latent space. In the anomaly localization case, residual image computation and activation map extraction techniques are used.

3.2.3. 2D axial patches

The 800 images are again split into a training set of 560 (80%), a validation set of 160 (20%), and a testing set of 80 (10%). For this experiment, patches are generated with a size of 32x32 pixels from each image. The way in which the patches are obtained differs from training and testing. In the training case, only patches having a center in the brain mask are obtained, i.e., no empty patches are visible during training. On the other hand, for testing, a sliding window approach is used to extract the patches with a stride of 16. The only-background-containing patches are ignored when applying anomaly detection and localization methods.

For the anomaly detection methods, the reconstruction error-based and the outlier detection in latent space-based methods are applied. Each patch is assigned one value referring to a normal or anomalous patch. After stacking the patches back together, the final image is a coarse segmentation mask with 16 pixels of resolution.

In the case of extracting an activation map and computing the residual image for anomaly segmentation, each pixel is assigned an activation value. Therefore, the final image is a summation of all pixel values with the original resolution. These results are compared with the segmentation results obtained in the 2D axial slices experiments.

3.2.4. Evaluation using toy anomalies

The evaluation of the anomaly detection and localization methods is conducted using toy anomalies generated in test images of the MOOD challenge dataset. The anomalies are generated by inserting a 2D shape (circle or random shape) in 50 slices of an image. 4 datasets are created:

- Anomaly dataset 1: circles with varying radius and intensity, which was similarly done in the MOOD challenge

[11] resulting in 20 different variations.

- Anomaly dataset 2: random shapes with varying size and intensity to check the impact of the shape in contrast to evaluation 1, which accounts for 20 combinations.
- Anomaly dataset 3: multiple random shapes with different combinations of size and intensity. The same size is kept while changing the intensity in 3 combinations and the same intensity is set with varying sizes in another 3 combinations.
- Anomaly dataset 4: black circular shape surrounded by a bright ring which is then blurred, simulating a tumor.

4. Results

4.1. Fashion MNIST

The mAP for anomaly detection using the reconstruction error in the Fashion MNIST dataset is 0.872. The AP values for the 9 anomalous items are depicted in Fig. 5. When the model is trained in t-shirts, the AP is lower for dresses (0.749) and shirts (0.666) than for ankle boots (0.997) and sandals (0.968).

Fig. 5: AP results for anomaly detection using reconstruction error in the 9 ”anomalous” items.

The individual AP values for each item using the different outlier detection techniques are displayed in Fig. 6. The highest mAP of the outlier detection methods is 0.894, corresponding to the covariance technique. In this case, dresses (0.909) and shirts (0.892) have similar results as for ankle boots (0.909) and sandals (0.895).

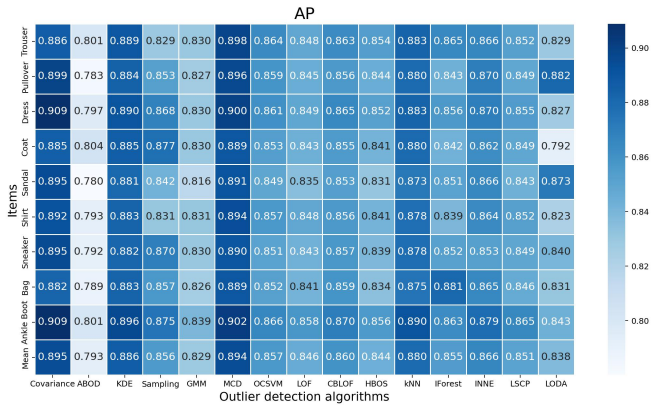

Fig. 6: AP results for anomaly detection using different methods for outlier detection in the 9 ”anomalous” items.

4.2. MOOD Challenge

4.2.1. 2D Axial slices

The mAP results for anomaly detection using the reconstruction error for circles, random shapes, multiple anomalies, and tumor-like structures are 0.4005, 0.5345, 0.3302, and 0.6844, respectively. The AP results for each combination are shown in Fig. 13 in Appendix B.1, which show higher values for darker, brighter and bigger anomalies.

The AP results for outlier detection in latent space for every analyzed technique are depicted in Figs. 15, 16, 17, 18, and 19 in Appendix B.2. The method with the highest mAP in each toy image dataset is OCSVM (0.5374), Covariance (0.5849), KNN (0.5741), ABOD (0.5655), and KDE (0.5387) for circles, random shapes, multiple anomalies with the same size, multiple anomalies with the same intensity, and tumor-like anomalies, respectively. The breakdown of the AP results is shown in Fig. 14 in Appendix B.2, which show no variation intensity and size-wise.

The mAP results for both anomaly detection methods for each toy image dataset are summarized in Fig. 7.

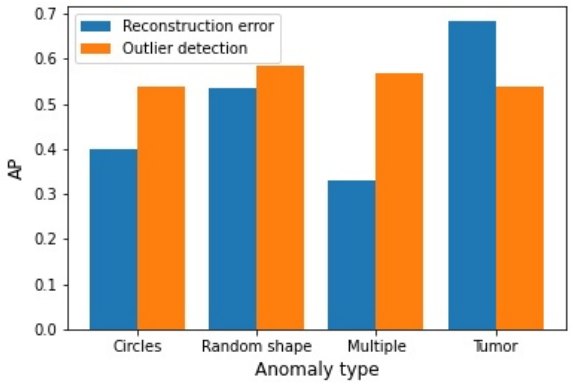

Fig. 7: mAP results for the two analyzed methods for anomaly detection (reconstruction error-based and outlier detection in latent space.

The mAP results for anomaly segmentation using the residual image computation in the four toy image datasets are 0.2978, 0.3179, 0.5159, and 0.5692 for circles, random shapes, multiple anomalies, and tumor-like structures, respectively. In the case of the extraction of activation maps, the mAP for circles, random shapes, multiple anomalies, and tumor-like structures are 0.4583, 0.5041, 0.6045, and 0.857, respectively. The breakdown of the AP values for each anomaly type for the two segmentation methods are displayed in Figs. 20 and 21 in Appendix B.3 and B.4, where a similar trend is observed with higher values for brighter, darker and bigger anomalies. A summary of anomaly segmentation using the residual image and activation map extraction is depicted in Fig. 8. An example of the segmentation methods applied to a toy image from each dataset is shown in Fig 9.

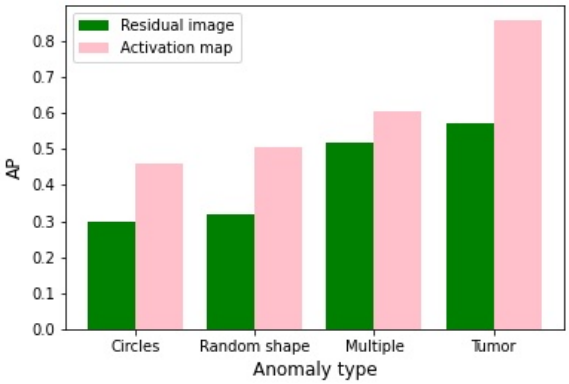

Fig. 8: mAP results for the two analyzed methods for anomaly segmentation (residual image computation and activation map extraction.

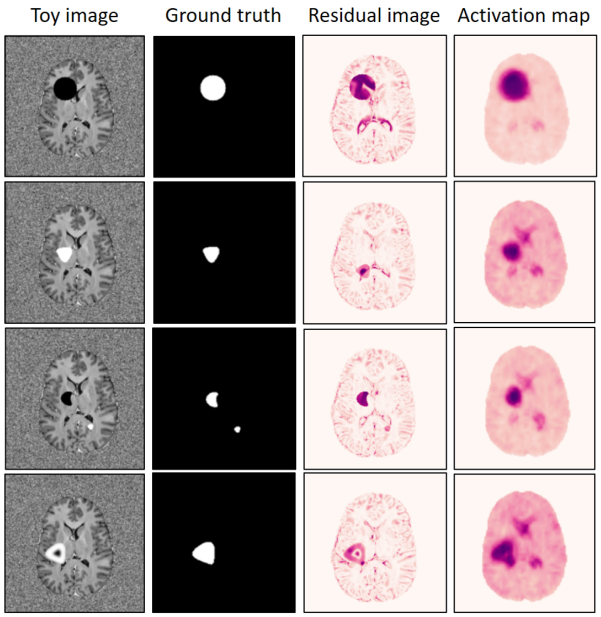

Fig. 9: Example of anomaly segmentation for toy slices containing (from top to bottom) a dark cylinder, a bright random shape, a combination of dark and bright anomalies, and a tumor-like structure.

4.2.2. 2D Axial patches

The mAP results for the reconstruction error-based technique applied to patches to localize anomalies are 0.1933, 0.2194, 0.2469, and 0.6942 for circles, random shapes, multiple anomalies, and tumor-like anomalies, respectively. In the case of residual error computation, the mAP are 0.05893, 0.0755, 0.0947, and 0.3074 for circles, random shapes, multiple anomalies, and tumor-like anomalies, respectively. Lastly, the activation map approach results in mAP of 0.2107, 0.2098, 0.3492, and 0.3614 for circles, random shapes, multiple anomalies, and tumor-like anomalies, respectively. These results are summarized in Fig. 10. It is important to note that the outlier detection methods are excluded from this analysis as the results were insignificant. The breakdown of the combination of the different anomalies is included in Figs. 22, 23 and 24 in Appendix C.1, C.2, and C.3, where again brighter, darker and bigger anomalies hvae higher AP values. Examples of the anomaly maps resulting from the 3 methods are shown in Fig. 11.

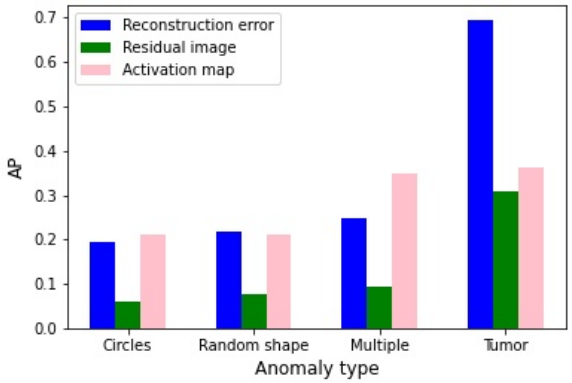

Fig. 10: Summary of the mean AP results of the anomaly segmentation methods using patches.

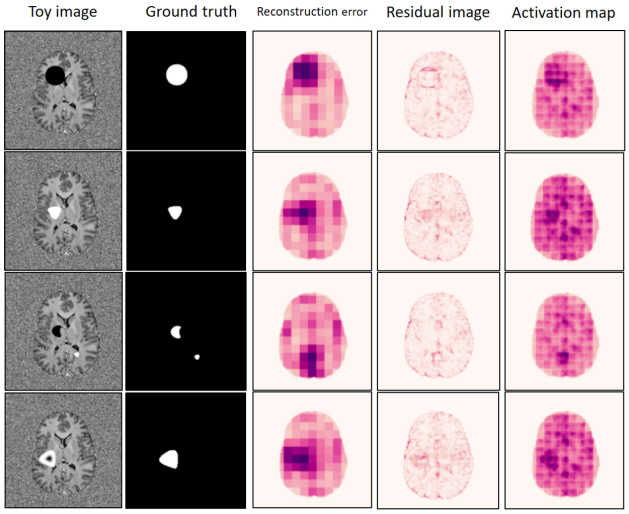

Fig. 11: Example of anomaly segmentation for toy slices containing (from top to bottom) a dark cylinder, a bright random shape, a combination of dark and bright anomalies, and a tumor-like structure.

5. Discussion

In this work, we propose the inspection of the latent space for anomaly detection, the introduction of activation maps for segmentation, and the incorporation of noise in the background

of the images to reduce the sensitivity to bright lesions. The methods are evaluated in 4 sets of toy images containing anomalies with different combinations of size, intensity, shape, and number.

Anomaly detection methods are first studied in the Fashion MNIST dataset, setting one of the items as normal and the rest as anomalous. The individual APs for each of the items show that items closer in latent space, with similar characteristics to the training images, are harder to distinguish. On the other hand, the individual AP scores in the case of the outlier detection in latent space show less variation in the similar and different items. Therefore, outlier detection in latent space proves to be a promising anomaly detection method, especially for those images that are more similar to the training set.

The anomaly detection problem is translated to the MOOD dataset, where the model is trained in healthy (normal) brain image slices. The results for anomaly detection show that the highest mAP of the outlier detection methods in latent space is greater than the mAP for the reconstruction error technique in all the toy anomaly datasets, except for the tumor-like structures. In the case of reconstruction error-based anomaly detection, the effect of intensity, size, shape, and number can be observed in the breakdown of the different anomaly combinations. These results show that bright and big anomalies are better detected, which was also discovered in [13], while the method fails for grey anomalies. Additionally, dark images show great results, similar to those from bright anomalies. These findings agree with the fact that the performance is lower for images with similar characteristics to those in the trained latent space. In this case, brain scans with small and grey anomalies are closer to healthy brain images.

Another reason to explain the poor performance of reconstruction error for grey anomalies is that the region of the anomaly in the normal counterpart generated by the decoder has a higher probability to have grey-level intensity pixels corresponding to the grey matter, which accounts for the majority of the brain (see Fig. 12). Therefore, the calculation of the reconstruction error will likely be low for images with grey anomalies.

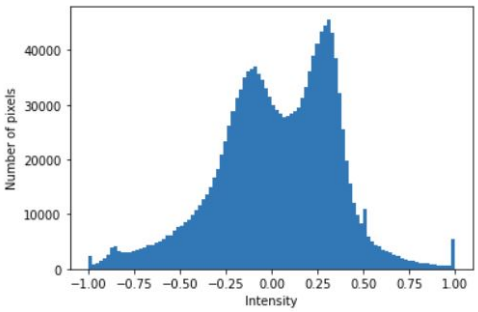

Fig. 12: Histogram of the brain region of a pre-processed T1-weighted brain MR image of a healthy patient from the MOOD dataset.

In contrast to the results for the reconstruction-based technique, the outlier detection methods in latent space show similar performance for the different combinations. Only 3 out of the 20 combinations in the circular anomalies show lower results, which correspond to brighter and darker big anomalies. In the random shapes, 16 of the combinations present higher AP values. Additionally, for the multiple anomalies, the results for outlier detection are higher in all cases. According to these results, outlier detection methods in latent space prove to have the potential for anomaly detection in brain MR images for images that are similar to the training data.

However, despite outlier detection methods having better mAP values, it should be taken into consideration that in the case of circles and random shapes, 60% of the test data contains grey anomalies, which is where the reconstruction error falters. For a dataset containing bright and dark anomalies, such as the toy images simulating tumors, the reconstruction error technique outperforms the analyzed outlier detection methods. Therefore, further investigation regarding the methods for outlier detection is needed. A future topic for research could be focused on the training of an outlier detection method and the model at the same time, similarly to what was done in [14].

Concerning anomaly segmentation, the method using the extraction of activation maps [15] is analyzed as a possible replacement for residual image computation. The results show that the mean AP value for activation maps is higher than for the residual images in all anomaly datasets. The intensity and size effect previously discussed in the reconstruction error method is also appreciated in the case of anomaly segmentation for both residual image computation and activation map extraction. For the latter technique, 18 out of 20 combinations of circular and random-shaped anomalies have better AP values than the residual image. This shows that using activation maps is a promising method for anomaly segmentation. However, they still struggle with more grey anomalies. Concerning the grey anomalies, a similar issue as for the reconstruction error can be found when using the residual image. That is, the region of the anomaly in the normal counterpart of the input image has a higher probability to have grey-value pixels. Therefore, when subtracting the input by its reconstruction, the error is smaller. This was also suggested in [13].

Finally, the results of applying the mentioned methods using a model trained in patches showed a lower performance for anomaly segmentation, except for the tumor-like structures, where the reconstruction-error-based technique achieves greater results. The effect of size and intensity can also be observed in the breakdown into the different combinations. However, the overall poor results of applying the anomaly scoring methods to the model trained with patches suggest that the model could be further improved.

It is important to mention that contrary to recent publications [11], [13], this work has proven to detect and segment dark anomalies along with bright lesions. This achievement confirms that changing the black background of brain MR images has an important effect in the detection and segmentation of anomalies depending on their intensity values. This is

also affirmed after the experiments using the model trained in patches with no background are able to localize dark lesions.

The evaluation proposed in this work suggests that anomaly detection and segmentation systems can be tested during development using a simple set of toy anomalies where the variations in shape, intensity, and size can easily be adapted for several analyses. However, for clinical practice, further investigation should be conducted. The following step in this process would consist of evaluating the techniques using real anomaly-containing data to analyze how the findings in the synthetic datasets translate to real scenarios.

6. Conclusion

This work tackles the current challenges of using VAEs for anomaly detection and segmentation of anomalies in brain MR images, which include the sensitivity towards bright anomalies and problems when using the reconstruction error or residual image for anomaly scoring. To that purpose, multiple anomaly detection and localization methods are compared, using toy images for their evaluation.

The results show that adding noise to the background of brain scans allows the identification of both dark and bright lesions, through the analyzed anomaly detection and segmentation systems. The analysis of outlier detection in latent space suggests that it has the potential to detect more similar anomalies to the training set, as demonstrated in the Fashion MNIST dataset. The extraction of activation maps proves to be a possible replacement for the computation of residual images for anomaly segmentation achieving better results in the evaluation of the model trained in brain MR slices. However, there is still room for improvement and further needed analyses for the system to be applied in real clinical scenarios.

References

[1] J. N. Itri, R. R. Tappouni, R. O. McEachern, A. J. Pesch, and S. H. Patel, "Fundamentals of diagnostic error in imaging," *RadioGraphics*, vol. 38, no. 6, pp. 1845–1865, Oct. 2018. [Online]. Available: https://doi.org/10.1148/rg.2018180021

[2] J. Latif, C. Xiao, A. Imran, and S. Tu, "Medical imaging using machine learning and deep learning algorithms: A review," in *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, 2019, pp. 1–5.

[3] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013. [Online]. Available: https://arxiv.org/abs/1312.6114

[4] C. Baur, B. Wiestler, S. Albarqouni, and N. Navab, "Deep autoencoding models for unsupervised anomaly segmentation in brain MR images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer International Publishing, 2019, pp. 161–169. [Online]. Available: https://doi.org/10.1007/978-3-030-11723-8_16

[5] D. Zimmerer, F. Isensee, J. Petersen, S. Kohl, and K. Maier-Hein, "Unsupervised anomaly localization using variational auto-encoders," in *Lecture Notes in Computer Science*. Springer International Publishing, 2019, pp. 289–297. [Online]. Available: https://doi.org/10.1007/978-3-030-32251-9_32

[6] D. Zimmerer, S. A. A. Kohl, J. Petersen, F. Isensee, and K. H. Maier-Hein, "Context-encoding variational autoencoder for unsupervised anomaly detection," 2018. [Online]. Available: https://arxiv.org/abs/1812.05941

[7] X. Chen and E. Konukoglu, "Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders," 2018. [Online]. Available: https://arxiv.org/abs/1806.04972

[8] S. You, K. C. Tezcan, X. Chen, and E. Konukoglu, "Unsupervised lesion detection via image restoration with a normative prior," in *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, ser. Proceedings of Machine Learning Research, M. J. Cardoso, A. Feragen, B. Glocker, E. Konukoglu, I. Oguz, G. Unal, and T. Vercauteren, Eds., vol. 102. PMLR, 08–10 Jul 2019, pp. 540–556. [Online]. Available: https://proceedings.mlr.press/v102/you19a.html

[9] W. H. Pinaya, P.-D. Tudosiu, R. Gray, G. Rees, P. Nachev, S. Ourselin, and M. J. Cardoso, "Unsupervised brain imaging 3d anomaly detection and segmentation with transformers," *Medical Image Analysis*, vol. 79, p. 102475, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841522001220

[10] C. Baur, S. Denner, B. Wiestler, N. Navab, and S. Albarqouni, "Autoencoders for unsupervised anomaly segmentation in brain MR images: A comparative study," *Medical Image Analysis*, vol. 69, p. 101952, Apr. 2021. [Online]. Available: https://doi.org/10.1016/j.media.2020.101952

[11] D. Zimmerer, J. Petersen, G. Köhler, P. Jäger, P. Full, K. Maier-Hein, T. Roß, T. Adler, A. Reinke, and L. Maier-Hein, "Medical out-of-distribution analysis challenge 2022," Mar. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6362313

[12] F. Meissen, G. Kaissis, and D. Rueckert, "Challenging current semi-supervised anomaly segmentation methods for brain mri," 2021. [Online]. Available: https://arxiv.org/abs/2109.06023

[13] F. Meissen, B. Wiestler, G. Kaissis, and D. Rueckert, "On the pitfalls of using the residual as anomaly score," in *Medical Imaging with Deep Learning*, 2022. [Online]. Available: https://openreview.net/forum?id=ZsoHLeupa1D

[14] Y. Zhang, X. Wang, Z. Ding, Y. Du, and Y. Xia, "Anomaly detection of sensor faults and extreme events based on support vector data description," *Structural Control and Health Monitoring*, vol. 29, no. 10, Jul. 2022. [Online]. Available: https://doi.org/10.1002/stc.3047

[15] J. Silva-Rodríguez, V. Naranjo, and J. Dolz, "Constrained unsupervised anomaly segmentation," *Medical Image Analysis*, vol. 80, p. 102526, Aug. 2022. [Online]. Available: https://doi.org/10.1016/j.media.2022.102526

[16] C. Doersch, "Tutorial on variational autoencoders," 2016. [Online]. Available: https://arxiv.org/abs/1606.05908

[17] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013. [Online]. Available: https://arxiv.org/abs/1312.6114

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] Y. Zhao, Z. Nasrullah, and Z. Li, "Pyod: A python toolbox for scalable outlier detection," *Journal of Machine Learning Research*, vol. 20, no. 96, pp. 1–7, 2019. [Online]. Available: http://jmlr.org/papers/v20/19-011.html

[20] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 233–240. [Online]. Available: https://doi.org/10.1145/1143844.1143874

[21] K. Boyd, K. H. Eng, and C. D. Page, "Area under the precision-recall curve: Point estimates and confidence intervals," in *Advanced Information Systems Engineering*. Springer Berlin Heidelberg, 2013, pp. 451–466. [Online]. Available: https://doi.org/10.1007/978-3-642-40994-3_29

[22] H. Xiao, K. Rasul, and R. Vollgraf. (2017) Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.

[23] T. Rohlfing, N. M. Zahr, E. V. Sullivan, and A. Pfefferbaum, "The SRI24 multichannel atlas of normal adult human brain structure," *Human Brain Mapping*, vol. 31, no. 5, pp. 798–819, Dec. 2009. [Online]. Available: https://doi.org/10.1002/hbm.20906

APPENDIX

A. Outlier detection algorithms

TABLE I: Outlier detection algorithms implemented in latent space for anomaly detection. These algorithms are obtained from Pyod.

Type	Name	Abbreviation
Covariance	Elliptic envelope	Covariance
Probabilistic	Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions	ECOD
Probabilistic	Angle-Based Outlier Detection	ABOD
Probabilistic	Kernel Density Functions	KDE
Probabilistic	Rapid distance-based outlier detection via sampling	Sampling
Probabilistic	Probabilistic Mixture Modeling	GMM
Linear model	Minimum Covariance Determinant	MCD
Linear model	One-Class Support Vector Machines	OCSVM
Linear model	Linear Method for Deviation-based Outlier Detection	LMDD
Proximity-Based	Local Outlier Factor	LOF
Proximity-Based	Cluster-based Local Outlier Factor	CBLOF
Proximity-Based	Histogram-based Outlier Score	HBOS
Proximity-Based	k Nearest Neighbors	KNN
Outlier Ensembles	Isolation Forest	IForest
Outlier Ensembles	Isolation-based Anomaly Detection Using Nearest-Neighbor Ensembles	INNE
Outlier Ensembles	Locally Selective Combination of Parallel Outlier Ensembles	LSCP
Outlier Ensembles	Lightweight On-line Detector of Anomalies	LODA

B. Extended results for the model using slices

B.1. Reconstruction error-based anomaly detection

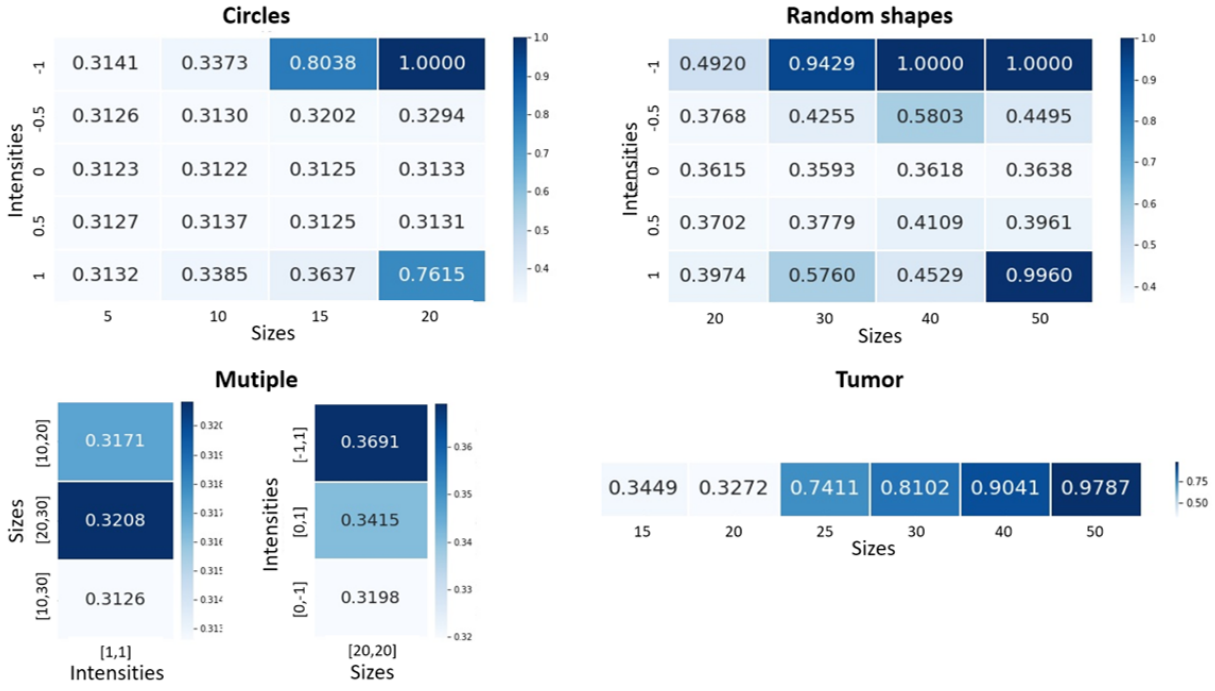

Fig. 13: A breakdown of the AP results for anomaly detection using the reconstruction error-based method in the different combinations of circles, random shapes, multiple anomalies, and tumor-like structures.

B.2. Outlier detection in latent space for anomaly detection

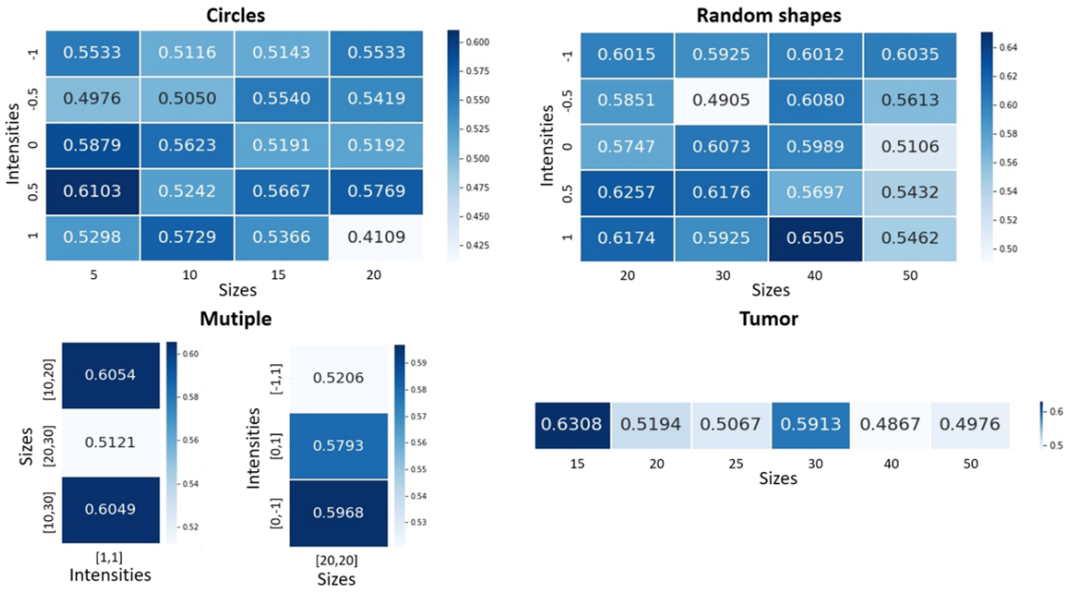

Fig. 14: A breakdown of the AP results for anomaly detection using the outlier detection in latent space technique in the different combinations of circles, random shapes, multiple anomalies, and tumor-like structures. The outlier detection method depicted for each anomaly dataset is the one with the highest mean AP score.

B.2.1. Circular anomalies

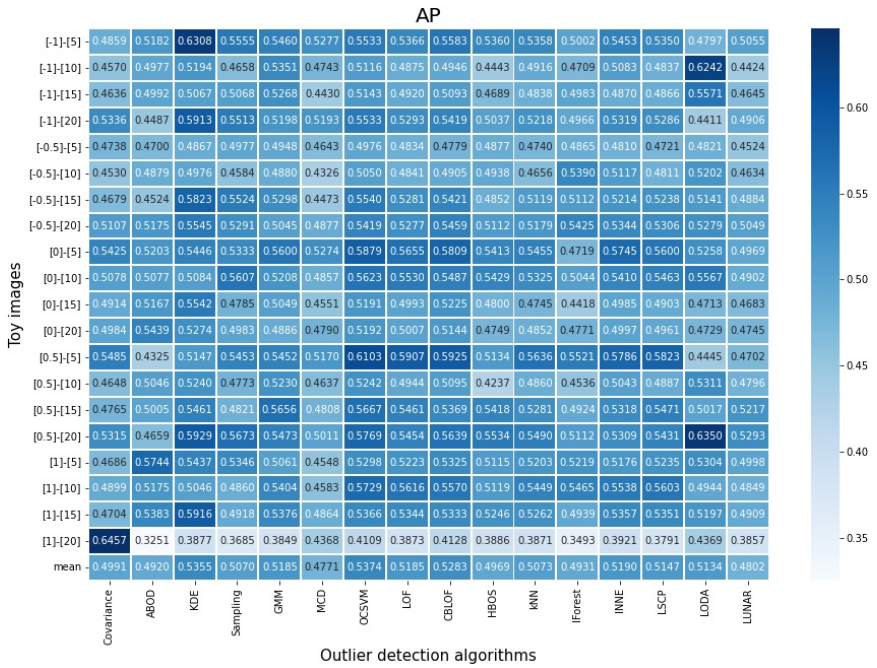

Fig. 15: AP results breakdown into the selection of outlier detection algorithms applied for the detection of circles. The highest mAP is ABOD.

B.2.2. Random shaped anomalies

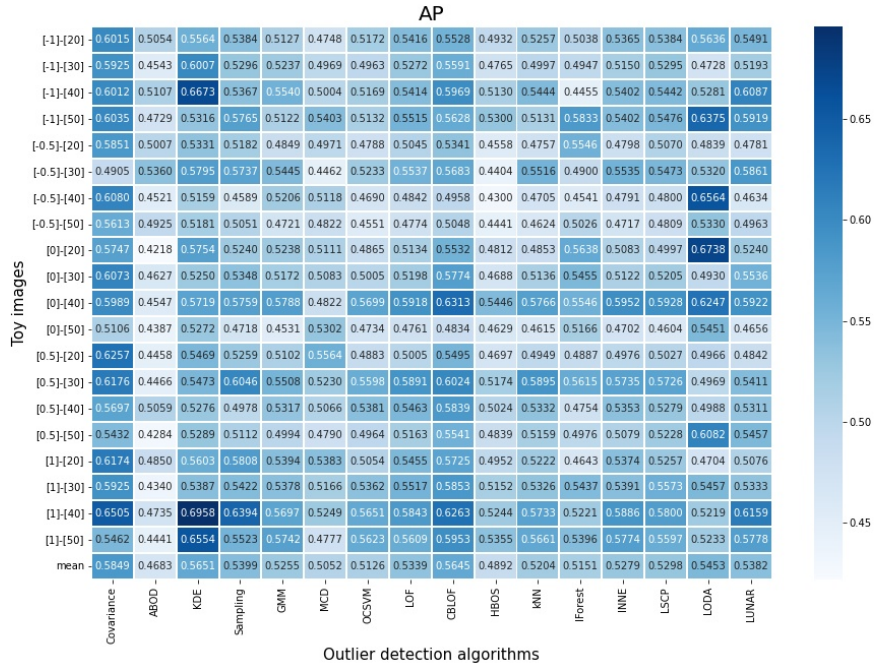

Fig. 16: AP results breakdown into the selection of outlier detection algorithms applied for the detection of random shapes.

B.2.3. Multiple anomalies

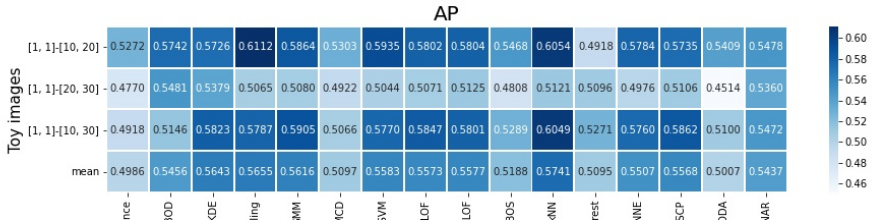

Fig. 17: AP results breakdown into the selection of outlier detection algorithms applied for the detection of multiple anomalies with the same intensity.

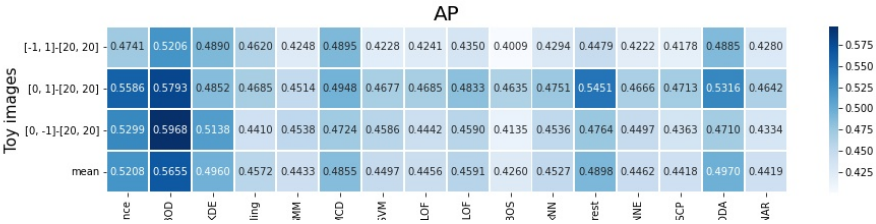

Fig. 18: AP results breakdown into the selection of outlier detection algorithms applied for the detection of multiple anomalies with the same size.

B.2.4. Tumor-like anomalies

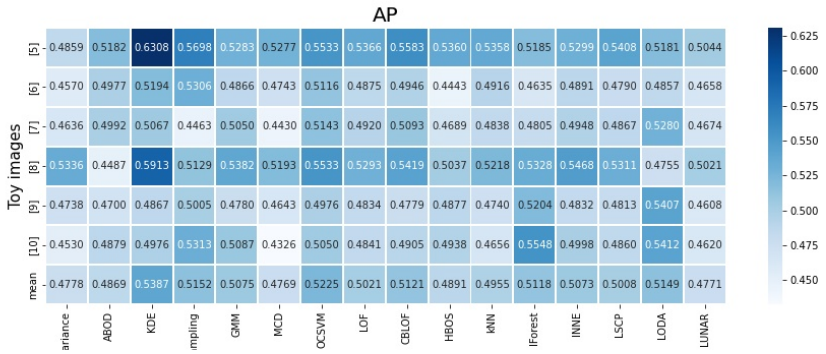

Fig. 19: AP results breakdown into the selection of outlier detection algorithms applied for the detection of tumor-like anomalies.

B.3. Residual image computation for anomaly localization

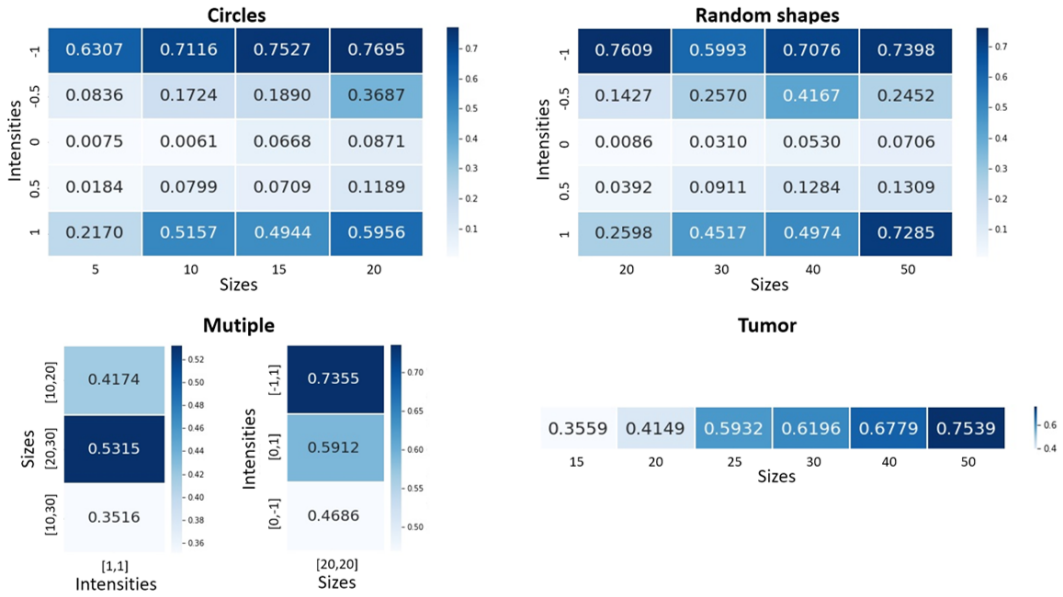

Fig. 20: A breakdown of the AP results for anomaly localization using the residual image computation in the different combinations of circles, random shapes, multiple anomalies, and tumor-like structures.

B.4. Activation map extraction for anomaly localization

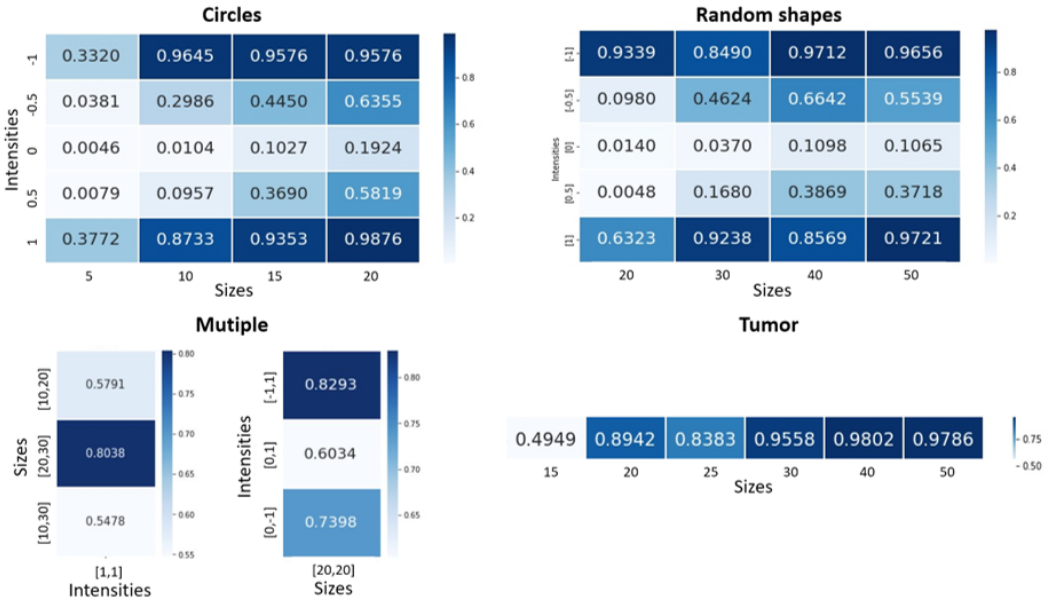

Fig. 21: A breakdown of the AP results for anomaly localization using the extraction of activation maps in the different combinations of circles, random shapes, multiple anomalies, and tumor-like structures.

C. Extended results for the model using patches

C.1. Reconstruction error

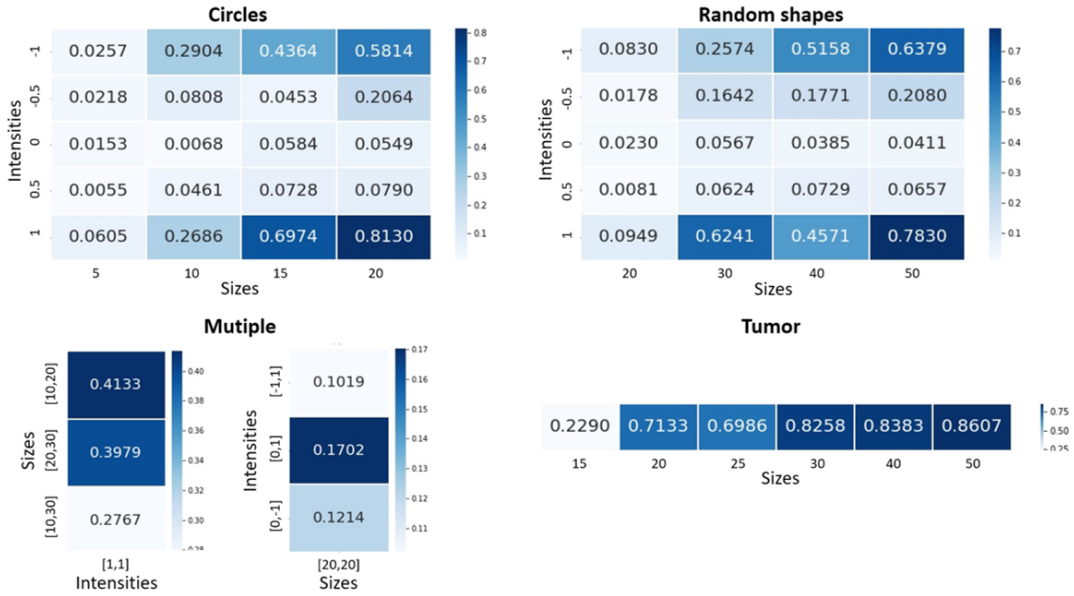

Fig. 22: A breakdown of the AP results for anomaly localization using the reconstruction error-based anomaly detection in patches in the different combinations of circles, random shapes, multiple anomalies, and tumor-like structures.

C.2. Residual image computation

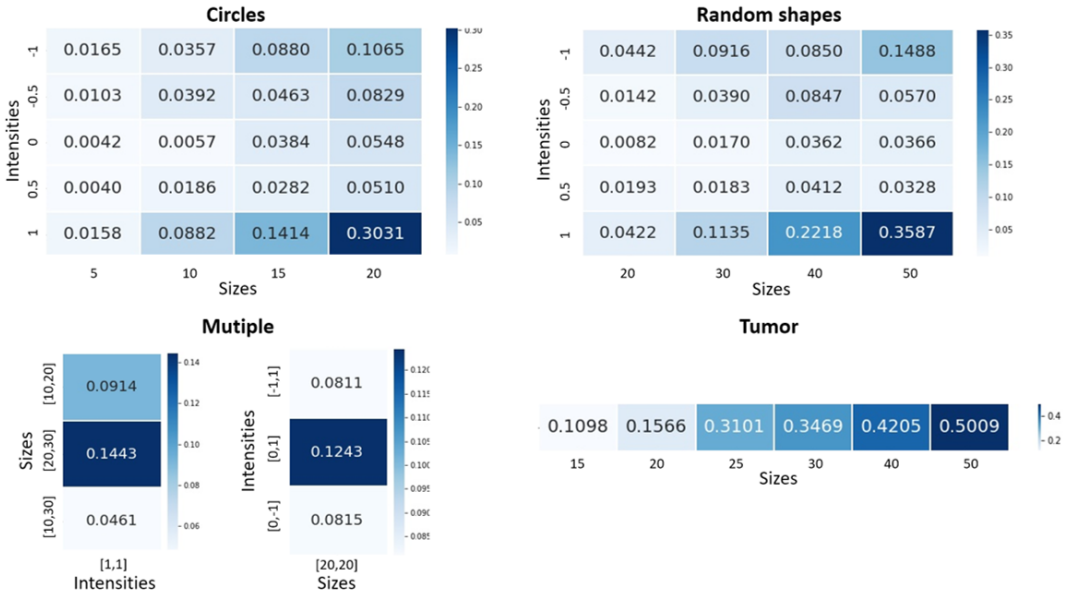

Fig. 23: A breakdown of the AP results for anomaly localization using the residual image computation in patches in the different combinations of circles, random shapes, multiple anomalies, and tumor-like structures.

C.3. Activation map extraction

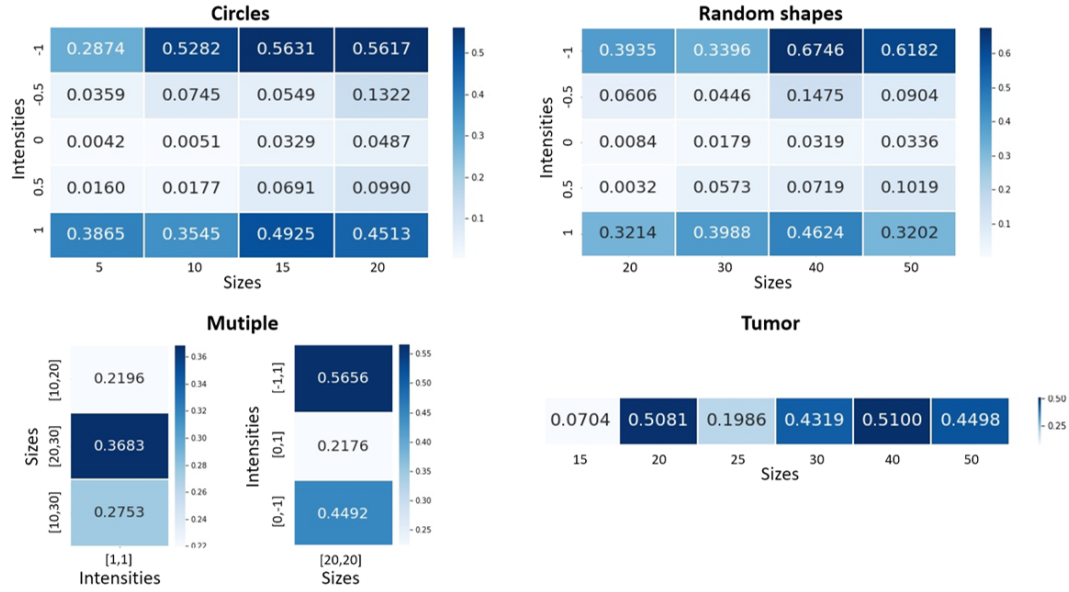

Fig. 24: A breakdown of the AP results for anomaly localization using the extraction of activation maps in patches in the different combinations of circles, random shapes, multiple anomalies, and tumor-like structures.