OCSVM-Guided Representation Learning for Unsupervised Anomaly Detection

Nicolas Pinon, Robin Trombetta and Carole Lartizien

***Abstract*—Unsupervised anomaly detection (UAD) aims to detect anomalies without labeled data, a necessity in many machine learning applications where anomalous samples are rare or not available. Most state-of-the-art methods fall into two categories: reconstruction-based approaches, which often reconstruct anomalies too well, and decoupled representation learning with density estimators, which can suffer from suboptimal feature spaces. While some recent methods attempt to couple feature learning and anomaly detection, they often rely on surrogate objectives, restrict kernel choices, or introduce approximations that limit their expressiveness and robustness. To address this challenge, we propose a novel method that couples representation learning with an analytically solvable One-Class SVM (OCSVM), through a custom loss formulation that directly aligns latent features with the OCSVM decision boundary. The model is evaluated on two tasks: a benchmark based on MNIST-C, and a challenging brain MRI lesion detection task. Unlike most methods that focus on large, hyperintense lesions at the image level, our approach succeeds to target small, non-hyperintense lesions, while we evaluate voxel-wise metrics, addressing a more clinically relevant scenario. Both experiments evaluate a form of robustness to domain shifts, including corruption types in MNIST-C and texture or population age variations in MRI. Results demonstrate performance and robustness of our proposed model, highlighting its potential for general UAD and real-world medical imaging applications. The source code is available at https://github.com/Nicolas-Pinon/uad_ocsvm_guided_repr_learning.**

***Index Terms*—Unsupervised anomaly detection, Representation learning, One-class SVM, Autoencoders, Joint optimization, MNIST-C, Medical imaging, Brain MRI**

I. Introduction

UNSUPERVISED anomaly detection (UAD) aims to identify patterns in data that deviate significantly from an underlying distribution learned from unlabeled normal samples. It is a critical problem in domains where anomalies are rare, variable, and costly to label, such as fraud detection or medical imaging. In neuroimaging, for instance, detecting subtle or small lesions in MRI scans without annotated anomalies remains an open challenge [1]. Models must not only detect rare and diverse outliers but also generalize reliably to new data distributions, such as those resulting from data acquired on different scanners, or populations with different demographics.

Existing methods fall into two main categories: reconstruction-based approaches and representation learning combined with support or density estimation methods. Autoencoders and their variants are frequently used in reconstruction-based strategies, under the assumption that anomalies will yield higher reconstruction errors. However, these models typically lack structured latent representations, which can lead to high quality reconstruction of never seen anomalies. To overcome this, other methods decouple representation learning from the anomaly scoring process, for instance by training a feature extractor independently from a classifier such as a one-class support vector machine (OCSVM [2]). However, this separation can yield to representations not optimized for the decision function computation, leading to suboptimal performance and limited generalization. Several recent approaches attempt to couple representation learning and anomaly detection more tightly, including methods inspired by Deep SVDD [3]. Yet, these methods often rely on approximations, suffer from hypersphere collapse, or impose strong inductive biases (e.g., linear kernel methods) that limit flexibility and robustness.

To address these limitations, we propose a novel method for UAD that tightly couples an autoencoder-based representation learning with a one-class SVM. Our core contribution lies in a new loss formulation that guides the encoder to produce latent representations optimized for the OCSVM's boundary. At each training step, the model splits latent samples into two subsets: one to fit the OCSVM boundary and another to enforce that new samples remain within it. This design reduces overfitting to non-relevant features by directly aligning the encoder's output with the SVM's discriminative objective. Crucially, it enables the use of an exact, analytically solved SVM objective, requiring no approximations or kernel restrictions, thereby preserving the full expressivity of the OCSVM.

To evaluate the proposed method, we conduct two experiments. First, we introduce a new benchmark task based on MNIST-C [4], a corrupted version of the MNIST dataset designed to simulate real-world anomalies. This task allows us to rigorously assess the model's performance in a controlled setting and compare it against state-of-the-art UAD methods. Importantly, this experiment evaluates the model's ability to perform anomaly detection under domain shift, as it must generalize across diverse corruption types. Second, we apply the model to two challenging medical imaging tasks: detecting large heterogeneous cancer lesions and detecting subtle (barely visible to the naked eye) brain lesions, both in MRI scans. In medical imaging, many UAD methods have traditionally focused on detecting large, hyperintense lesions, which are often visible and easy to identify, especially through reconstruction-based methods. Our work, in comparison, tackles the problem of detecting lesions that can be small and not necessarily hyperintense, representing a more subtle and clinically sig-

 (Corresponding author: Nicolas Pinon)

Nicolas Pinon, Robin Trombetta and Carole Lartizien are with Univ. Lyon, CNRS UMR 5220, Inserm U1294, INSA Lyon, UCBL, CREATIS, France (e-mails: nicolas.e.pinon@laposte.net; carole.lartizien@creatis.insa-lyon.fr ; robin.trombetta@creatis.insa-lyon.fr)

nificant problem. Additionally, while a significant portion of UAD studies in medical imaging measure performances at the image level, we also assess voxel-wise anomaly detection, thus evaluating precise localization of anomalies within the image. Furthermore, this second series of experiments in neuroimaging evaluates the model's robustness to domain shifts arising from variations in MRI scanners and patient demographics, such as age.

The contributions of this work are twofold:

- **A novel OCSVM-guided representation learning method for general UAD is introduced, based on a novel loss term aiming at optimizing the representation learner to produce more suitable representations when used in conjunction with OCSVM**
- **We demonstrate the method's applicability on the corrupted MNIST dataset under different domain shifts, as well as on real-world medical imaging tasks, showing improved sensitivity to subtle and non-hyperintense lesions in public brain MRI datasets.**

The remainder of this paper is organized as follows: section II reviews related work on anomaly detection, and then specifically methods used in medical imaging. Section III describes our proposed method, detailing the OCSVM-guided representation learning strategy. Sections IV and V present our experimental studies: digit distinction under corruptions using the MNIST-C dataset and lesion detection in brain MRI, respectively. Section VI provides a general discussion, including an analysis of the loss components and concludes the paper while outlining potential future research directions.

II. Related works

Unsupervised Anomaly Detection (UAD) methods can be broadly categorized into three main families: *reconstruction-based methods*, *density estimation-based methods*, and *support estimation-based methods*, as outlined in the review by Ruff et al. [5]. All methods share a common objective: modeling the distribution of normal (i.e., non-anomalous) data, often referred to as the normative distribution. Once this distribution is learned, anomalies can be detected as samples that significantly deviate from it.

Representation learning lies at the core of most unsupervised anomaly detection approaches. Reconstruction-based methods use representations as a compression/decompression mechanism, while in practice, both density- and support-based methods typically do not operate directly on raw data, but instead leverage intermediate feature representations, often learned through neural networks, to better capture the structure of normal data.

In the following bibliographic review, we specifically focus on autoencoders due to their simplicity and widespread use as a foundational method for unsupervised feature learning. While other feature extractors, such as transformer-based models, could also be employed in a similar framework, exploring all possible alternatives is beyond the scope of this work. We also place a slight emphasis on support estimation methods in order to highlight how our contribution compares to other methods in the same family.

Section II-A covers reconstruction-based methods where the anomaly score is directly derived from the reconstruction error. Section II-B focuses on support and density estimation methods that use learned representations. We distinguish between decoupled methods, where the representation learning and the anomaly scoring are optimized separately, and coupled methods, which jointly optimize both components, like the method proposed in this work. Finally, section II-C provides an overview of anomaly detection methods specifically applied to medical imaging.

A. Reconstruction-based methods

A widely used approach in UAD is to leverage an autoencoder's (AE) ability to reconstruct normal data while failing to accurately reconstruct anomalies. As detailed in the review by Ruff et al. [5], reconstruction-based methods assume that, after training on normal samples, an autoencoder will learn a compressed representation that captures essential features of the normal data distribution. When presented with an anomalous input, the reconstruction error is expected to be significantly higher due to the model's inability to generalize to unseen, out-of-distribution patterns.

Early approaches relied on simple autoencoders trained with standard mean squared error or cross-entropy loss, where anomalies were detected based on high reconstruction error [6]. This paradigm has been widely applied to image anomaly detection [7], [8], [9] and extended to various domains, such as industrial defect detection [10] or medical images [11]. Variational Autoencoders (VAE) introduced a probabilistic constraint on the latent space, which helps regularize representations, but they often struggle to clearly separate normal from anomalous reconstructions due to their tendency to generate blurry outputs [12].

Hybrid methods, known as restoration methods have emerged, which combine the reconstruction error with an estimation of the density of the distribution of normal samples in the autoencoder's latent space. These methods aim to "heal" the image by restoring it to the normal distribution (thus erasing the anomaly) and then comparing it to the original image through the reconstruction error. One example is the work by Wang et al. [13], which applies this approach to industrial images by using a quantized autoencoder (VQ-VAE) in conjunction with an autoregressive model (PixelSnail [14]) for density estimation in the latent space. Another type of restoration methods has gained recent popularity for anomaly detection in images: diffusion models, where the image is first partially noised, and then denoised with a UNet-like model, effectively providing a restored image [15].

Another alternative direction involves synthetic anomaly detection (also called self-supervised learning strategies [16], [17]), where synthetic anomalies are added to the data during training of a supervised method. This approach, also proved effective in medical imaging [18], [19], suffers from a severe drawback : the synthetic anomalies distribution must match the (unknown) true anomaly distribution, therefore imposing a strong prior on anomalies that can be detected.

Despite their effectiveness, Ruff et al. [5] highlight several limitations of reconstruction-based methods. Autoencoders

may generalize too well, inadvertently reconstructing anomalies with low error, which weakens their discriminative power [20]. Also, reconstruction error alone does not explicitly define a geometrically-coherent decision boundary between normal and anomalous data, making it hard to calibrate anomaly scores. These challenges motivate alternative approaches where autoencoders serve as representation learners rather than direct anomaly detectors, as discussed in section II-B.

B. Support/density estimation methods

Density and support estimation methods attempt to explicitly characterize the distribution of normal data either by modeling its density or by learning a decision boundary that encloses the normal data. They typically rely on representation learning techniques to effectively model the structure of normal data that can be coupled with classical methods like One-Class SVM (OCSVM [2]), Support Vector Data Description (SVDD [21] and their variants, Gaussian Mixture models, etc.

In this section, we distinguish between decoupled methods, where the representation learner is trained separately (II-B1) before applying a support or density estimation method, and coupled methods (II-B2), where the representation learning process is influenced by the anomaly detection objective.

1) Decoupled methods: A common approach is to first train an autoencoder to reconstruct its input, thus providing an encoder capable of producing a compressed representation of the input and then apply a separate anomaly detection method on the learned latent representations such as multivariate Gaussian in PaDiM [22], clustering in Perera and Patel [23] or OCSVM in [24], [8], [9] .

This OCSVM-based approach was applied to industrial images [9] and synthetic aperture radar images [24]. In both cases, a convolutional autoencoder is trained on normal samples, and the encoder's latent features are fed to an OCSVM for anomaly detection. In [24], the features are further reduced via PCA, and as in [8] a discriminator is used.

Decoupled methods often suffer from a sub-optimal alignment between the learned representations and the anomaly detection objective. This mismatch can lead to degraded performance, particularly in complex or high-dimensional settings where anomaly structures are subtle.

2) Coupled methods: Coupled methods aim to address this limitation by integrating the representation learning and support/density estimation steps into a unified framework, thereby encouraging the latent space to be more directly optimized for the detection task. A foundational example is Deep SVDD (DSVDD [3]), which replaces the implicit dual space mapping of traditional SVDD by an explicit modeling (thus approximated) with a neural network. The normal data points are projected in a dual space where they must fit into an hypersphere of learned radius (soft-variant) or just compacted around a predefined center (hard-variant). Anomalies are then identified by measuring the distance to the center (hard) or to the hypersphere (soft). The method is evaluated on several standard image datasets, including MNIST and CIFAR-10, where it demonstrates better performance than kernel-based baselines such as OCSVM.

Different approaches, namely DSPSVDD [25], DVAESVDD [26], DASVDD [27], CDSVDD [28] or Patch SVDD [29], are built on the seminal Deep SVDD formulation to both address the hypersphere collapse issue and attempt to improve the discriminative power of the learned representation. They all were shown to outperform it based on datasets such as MNIST, Fashion-MNIST [30] or MVTecAD [10].

Deep Structure Preservation SVDD (DSPSVDD) [25] enhances Deep SVDD by first pre-training an autoencoder and then adding the deep SVDD term in the loss for further fine-tuning. The major difference is that the reconstruction loss term is still present in the fine-tuning.

In a similar fashion, VAE-based Deep SVDD (DVAESVDD [26]) optimizes a VAE's reconstruction loss and the SVDD's hypersphere loss. Contrastive Deep SVDD (CDSVDD [28]) leverages contrastive learning to improve the discriminative power of the learned representations by minimizing both the contrastive loss and the SVDD loss.

Beyond SVDD-based formulations, Zong et al. [31] introduce the Deep Autoencoding Gaussian Mixture Model, which combines a compression network with a GMM applied in the latent space. The loss function integrates the reconstruction error, the GMM log-likelihood, and a regularization term; it was originally tested on tabular datasets (KDDCup99, Thyroid, Arrhythmia).

Other coupled methods include one-class GAN (OCGAN) [32], which uses adversarial training to enforce that every normal samples are distributed as a uniform distribution and that every interpolated sample from this distribution output a normal-looking image.

Overall, coupled methods seem to benefit from end-to-end optimization, where the representation learning and anomaly detection objectives are jointly optimized. This could ensure that the learned features are directly tailored for anomaly discrimination. While coupled approaches seem to surpass their decoupled counterpart in the cited studies, the diversity of evaluation protocols and datasets makes generalization of conclusions difficult. To the best of our knowledge, no comprehensive study has been conducted to systematically assess the benefits of coupling representation learning with anomaly detection, compared to decoupled approaches. Also, to this day, no method makes use of the full flexibility offered by the kernel-representation of OCSVM or SVDD: all methods use approximations or limitations regarding the type of kernel used for dual space mapping.

Moreover, most existing studies focus on standard, low-complexity datasets such as MNIST, Fashion-MNIST, or CIFAR-10, which do not reflect the challenges of real-world applications. In particular, the medical imaging domain, despite its complexity and practical importance, remains largely unexplored in this context. This highlights the need for a dedicated review of UAD methods in medical imaging, which we present in section II-C.

C. Unsupervised anomaly detection for medical images

In this section, we focus on Unsupervised Anomaly Detection (UAD) methods specifically applied to medical imaging.

While the broader field of medical anomaly detection encompasses a wide range of modalities and anatomical regions, we restrict our discussion to studies that align with our focus on brain MRI, based on recent reviews and benchmark analyses in the domain [17], [33].

Reconstruction-based methods presented in section II-A, have been widely applied to brain lesion detection in MRI. Baur et al. [11] conducted a comprehensive comparative benchmark of various autoencoder architectures, including classical, variational and adversarial autoencoders for detecting large tumor lesions (Gliomas) as well as hyperintense lesions in brain MRI datasets such as MSLUB [34] and MSSEG [35]. Their findings highlight the effectiveness of reconstruction-based approaches for identifying small, hyperintense lesions, which are common in conditions like multiple sclerosis. The same authors proposed a UNet-like autoencoder architecture [36] demonstrating strong performance for detecting small hyperintense lesions on the WMH challenge dataset.

Pinaya et al. [37] introduced a restoration-based approach using a Vector Quantized Variational Autoencoder (VQ-VAE) coupled with a transformer model for density estimation in the latent space. This method was evaluated on multiple neuroimaging datasets, including MSLUB, BraTS, and WMH, further highlighting the utility of reconstruction-based techniques for hyperintense lesion detection. Additionally, Ramirez et al. [38] used VAEs to detect anomalies in Parkinson's patients' brain MRI, while Zimmerer et al. [39] and Zhao et al. [40] employed VAEs for brain tumor segmentation, leveraging reconstruction errors as anomaly scores.

Diffusion models have been recently proposed as an alternative to auto-encoder based architectures in UAD for brain lesions detection in MRI. As for AE-based models, the underlying assumption is that these models will generate anomaly-free images when inputted pathological images at inference. The anomaly score map is then computed based on scoring functions accounting for the standard reconstruction error map. Like autoencoders, these models suffer from the need to compromise the sensitivity to anomalous regions, increasing with high corruption (noise for diffusion models and compression for AE), and the fidelity of the reconstructed normal regions (decreasing with noise/compression), as seen in [41]. Several architectures have been proposed to enhance detection performance of these models. Some authors proposed employing more elaborate noise types, e.g. multiscale or simplex noise in DAE [42] and AnoDDPM [43]) instead of standard Gaussian noise. Very recent contributions leverage information from the original image, either by conditioning the denoising process to the original images, like cDDPM [44], [45] or by masking strategies, to focus the denoising process on the lesion area so as to preserve the details of regions predicted as healthy, like AutoDDPM [46] and THOR [47] which apply implicit guidance only during inference or MAD-AD [48] which incorporates information from the original image during both training and inference.

In addition to reconstruction-based methods or methods based on synthetic anomaly generation, support/density estimation approaches, presented in section II-B, have also been applied to medical imaging. For example, we proposed to employ autoencoders as feature extractors, followed by OCSVM for anomaly detection [49], [50], [51]. In [49], we utilized a localized OCSVM approach to detect challenging epileptogenic lesions in a private dataset, while in [50] and [51], we proposed a patient-specific OCSVM framework evaluated on the WMH dataset in the former and on the public PPMI dataset (related to Parkinson disease) in the later. Furthermore, Azami et al. [52] and Bowles et al. [53] used OCSVM for brain MRI anomaly detection, the latter applying it to unsupervised brain lesion segmentation by modeling white and gray matter voxels.

Cai et al. [17] performed a wide benchmark on image-level anomaly detection on medical imaging datasets, and image-level and voxel-level anomaly detection specifically for brain MRI on different datasets including the BraTS dataset [54] depicting brain tumors but not any dataset depicting smaller brain lesions like WMH or MSLUB. While this benchmark does not evaluate support/density estimation methods on the voxel-level anomaly detection task, they evaluate a wide variety of methods based on reconstruction error including DAE [55], AnoDDPM [43] and AutoDDPM [46] or models based on synthetic anomaly generation. They find that reconstruction-based methods outperform other methods for voxel-level anomaly detection. On the BraTS dataset, DAE [55] is shown to outperform autoDDPM [46], itself outperforming anoDDPM with simplex noise [43].

As pointed out in the literature [33], recent benchmark studies dedicated to the detection and localization of small lesions (e.g. MSLUB or WMH) are lacking. Results compiled from the 2021 benchmark study of Baur et al [11] for autoencoder based models and from recent papers including state-of-the art diffusion models ([48], [44]) highlight that cDDPM [44], [45] combined with an anomaly scoring function based on the Mahalanobis distance outperformed anoDDPM [43], which itself performed better than DAE [55] for the detection and localisation of WMH and MSLUB lesions. In two recent studies of Beizaee et al. [48], [56], anoDDPM [43] was also shown to outperform autoDDPM [46], itself performing better than DAE [55] on small stroke lesions of the ATLAS dataset.

The recent review of Behrendt et al [33] pointed out another issue in the evaluation of medical anomaly detection methods in brain MRI related to the predominance of hyperintense lesions in benchmark datasets. As first noted by Meissen et al. [57], many state-of-the-art methods are evaluated on anomalies that are significantly brighter than the surrounding tissue in the MRI image (e.g. FLAIR), such as those in the BraTS and WMH datasets. This raises concerns about the generalizability of these methods to more challenging anomalies, such as those with subtle intensity differences or complex morphological characteristics. Some groups including ours ([58], [59]) demonstrated that simply thresholding these MRI images could achieve competitive performance on hyperintense lesion detection e.g. WMH lesion in FLAIR imaging, highlighting the need for more rigorous evaluation protocols and diverse datasets.

Despite encouraging results on hyperintense lesions, the performance of unsupervised anomaly detection methods on more challenging, publicly available medical imaging datasets

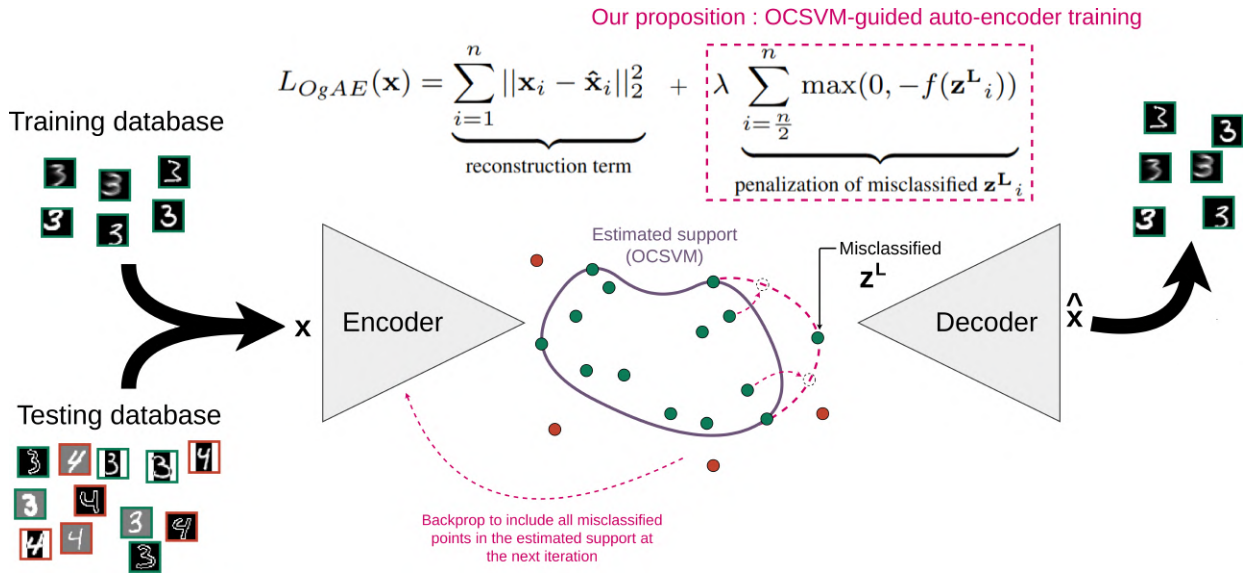

Fig. 1: Graphical abstract of the proposed method. During training, the autoencoder must both minimize the reconstruction error between input and output and a new loss term (section III-B) that guides the encoder towards representations that are more fitted for support estimation with OCSVM.

remains largely unevaluated. Autoencoder-based reconstruction methods continue to serve as strong baselines. In contrast, support and density estimation approaches (decoupled II-B1) remain underexplored in this context, often evaluated only on private datasets or omitted from comparative benchmarks. Also, to the best of our knowledge, no coupled (II-B2) method that jointly optimizes feature representation and anomaly detection has been applied to medical imaging.

III. Method: OCSVM-guided representation learning

The method we propose is presented in Figure 1. An autoencoder is used for representation learning, while a OCSVM is used to estimate the normal data distribution support. The main goal of the term that we add to the loss function of the autoencoder is to use normal samples that are misclassified during training (projected outside the support) to modify the representation space such that these misclassified samples will be included in the estimated support at the next iteration. As we stated in section II-B, the proposed architecture is compatible with any representation learner (including transformers). Section III-A details the main idea of the method without coupling by describing the representation learning step III-A1, followed by the anomaly detection step III-A2. Section III-B describes our contribution: coupling of the two steps through the OCSVM-guidance of the representation learning.

A. Decoupled representation learning and anomaly detection

1) Representation learning with autoencoder: To learn efficient and compressed representations, we train the autoencoder to reconstruct as accurately as possible an input batch of normal samples [1] $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, based on the classical MSE loss:

$$L_{AE}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2 \tag{1}$$

Where $\hat{\mathbf{x}}_i$ is the reconstruction of $\mathbf{x}_i$. After training, the decoder is discarded and the encoder is used, frozen, to perform dimensionality reduction of samples $\mathbf{x}$ into their latent representation $\mathbf{z}$.

2) Anomaly detection with one-class SVM: To perform the detection of anomalies, we estimate the support of the normal data (the boundaries of the normative distribution) with a One-Class SVM (OCSVM [2]). This is done by constructing a decision function f, positive on the estimated support of the distribution of normal samples $\mathbf{z}_i$, negative elsewhere and null on the frontier. The normal samples are first mapped to a high dimensional space by a feature map $\mathbf{\Phi}(\cdot)$ associated with a kernel k such that $k(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{\Phi}(\mathbf{z}_i) \cdot \mathbf{\Phi}(\mathbf{z}_j)$. As the problem is linear in this re-description space, the parameters $\mathbf{w}$ and ρ of the hyperplane $\mathbf{w} \cdot \mathbf{\Phi}(\mathbf{z}) - \rho = 0$ are obtained by solving a convex optimization problem, presented in equation 2, aiming at maximizing the distance of the hyperplane from the origin.

$$\begin{aligned} \min_{\mathbf{w},\rho,\boldsymbol{\xi}} \quad & \frac{1}{2}||\mathbf{w}||^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\xi_i - \rho \\ \text{subject to} \quad & \langle \mathbf{w}, \mathbf{\Phi}(\mathbf{z}_i)\rangle \geq \rho - \xi_i \qquad i \in [1, n] \\ & \xi_i \geq 0 \qquad i \in [1, n] \end{aligned} \tag{2}$$

The decision function can then be expressed as $f(\mathbf{z}) = \mathbf{w}^* \cdot \mathbf{\Phi}(\mathbf{z}) - \rho^*$, with $\mathbf{w}^*$ and ρ^* the solutions of the optimization problem.

Through a process known as the kernel trick, the problem is actually solved in its dual form leading to the following expression of the decision function

$$f(\mathbf{z}) = \sum_{j=1}^{n} \alpha_j^* k(\mathbf{z}_j, \mathbf{z}) - \rho^* \tag{3}$$

which corresponds to a weighted mean of the kernel distance to each normal samples $k(\mathbf{z}_j, \mathbf{z})$. The Lagrange multi-

[1] Input batches in experiment 1 will be batches of whole images and in experiment 2 batches of image patches, but this method can be used with any type of data (even non-image, if the autoencoder is adapted).

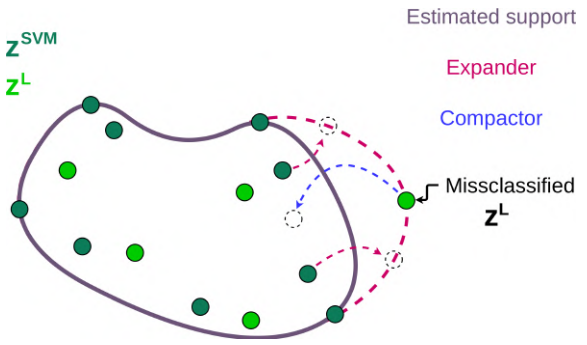

Fig. 2: Visualization of the two terms present in the proposed loss : both terms are based on the idea to use the misclassified $\mathbf{z}^{\mathbf{L}}$ to steer the representations towards SVM-compatible features. While the expander term focus on moving the $\mathbf{z}^{\mathbf{SVM}}$ to expand the estimated support, the compactor term focus on moving the $\mathbf{z}^{\mathbf{L}}$ inside the estimated support.

pliers of the dual problem (α_j^*) are sparse and ρ^* is derived from them.

At inference, to obtain the anomaly score of a new sample $\mathbf{x}$, it must first go through the encoder to obtain its latent representation $\mathbf{z}$, and then through the decision function f. Note that this score will be positive if the sample is within the distribution and negative if outside. The more negative the score, the further the sample is from the normal distribution and thus the more suspicious it will be considered.

B. Coupling: OCSVM-guidance of the representation learning

We describe in this section our contribution: a novel OCSVM-guidance (Og) loss term. The goal of this loss is to align as best as possible the representation of the encoder with the downstream task of estimating the support of the normal distribution with the OCSVM. This is performed by splitting each training batch into two, one used for support estimation ($\mathbf{z}^{\mathbf{SVM}}$) and one for loss computation ($\mathbf{z}^{\mathbf{L}}$).

The OCSVM-guidance is divided into two terms: the expander and the compactor, as represented in Figure 2. The compactor term makes the estimated support more compact by *moving misclassified normal training samples* inside the estimated support: this ensures the support stays compact and allows anomalies to fall outside the support. To prevent collapsing of the support, as can happen in deep SVDD, the expander term *moves the boundary* such that misclassified normal training samples fall inside the estimated support. By training the encoder to align with the estimated support (whether by expanding it or compacting it), we implicitly encourage deviations from this manifold to correspond to anomalous behavior, thus learning OCSVM-compatible features while avoiding irrelevant overfitting.

As stated, one batch of samples, after encoding, $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ is split into two: the part used to solve the OCSVM problem ($\mathbf{z}^{\mathbf{SVM}}$) and the other for the loss computation ($\mathbf{z}^{\mathbf{L}}$):

$$\mathbf{z}_i = \begin{cases} \mathbf{z}_i^{SVM} & \text{for } 1 \leq i \leq \frac{n}{2}, \\ \mathbf{z}_i^{L} & \text{for } \frac{n}{2} < i \leq n. \end{cases} \tag{4}$$

At each batch, we solve the OCSVM problem for the ${\mathbf{z}^{\mathbf{SVM}}}_i$, which will give the optimal $\boldsymbol{\alpha}$ and ρ: $\boldsymbol{\alpha}^*$ and ρ^*.

The proposed L_{OgAE} loss is composed of a standard reconstruction error term and the OCSVM-guidance (Og) term, which penalizes the misclassified ${\mathbf{z}^{\mathbf{L}}}_i$ (that are not used to compute the SVM problem):

$$L_{OgAE}(\mathbf{x}) = \underbrace{\sum_{i=1}^{n} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2}_{\text{reconstruction term}} + \lambda \underbrace{\sum_{i=\frac{n}{2}}^{n} \max(0, -f({\mathbf{z}^{\mathbf{L}}}_i))}_{\text{penalization of misclassified } {\mathbf{z}^{\mathbf{L}}}_i} \tag{5}$$

The second term, weighted by λ, indeed penalizes only the misclassified ${\mathbf{z}^{\mathbf{L}}}_i$, as the decision function outputs positive values for correctly classified ${\mathbf{z}^{\mathbf{L}}}_i$, and thus $\max(0, -f({\mathbf{z}^{\mathbf{L}}}_i))$ is 0. Misclassified ${\mathbf{z}^{\mathbf{L}}}_i$ are penalized proportionally to their euclidean distance to the estimated hyperplane.

The interest of separating the latent representation vectors into two parts $\mathbf{z}^{\mathbf{SVM}}$ and $\mathbf{z}^{\mathbf{L}}$ appears here: as the SVM frontier is estimated on the ${\mathbf{z}^{\mathbf{SVM}}}_i$, most of them are correctly classified. This justifies the use of another set of latent vectors $\mathbf{z}^{\mathbf{L}}$. Penalizing samples not used for the support estimation could also be viewed as a way to penalize bad generalization to unseen samples. We can develop L_{OgAE} with the expression of f from equation 3:

$$\begin{aligned} L_{OgAE}(\mathbf{x}) = &\sum_{i=1}^{n} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2 \\ &+ \lambda \sum_{i=\frac{n}{2}}^{n} \max\left(0, -\sum_{j=1}^{\frac{n}{2}} \alpha_j^* k({\mathbf{z}^{\mathbf{SVM}}}_j, {\mathbf{z}^{\mathbf{L}}}_i) - \rho^*\right) \end{aligned} \tag{6}$$

Recall that $\boldsymbol{\alpha}^*$ and ρ^* are functions of the ${\mathbf{z}^{\mathbf{SVM}}}_i$. If we separate the λ-term into what depends on the ${\mathbf{z}^{\mathbf{SVM}}}_i$ and what depends on the ${\mathbf{z}^{\mathbf{L}}}_i$, using the stopgradient operator sg[.] and two linked coefficients $\beta_1 + \beta_2 = 1$, we can write L_{OgAE} as:

$$\begin{aligned} &L_{OgAE}(\mathbf{x}) = \textstyle\sum_{i=1}^{n} ||\mathbf{x}_i - \hat{\mathbf{x}}_i||_2^2 \\ &\overbrace{+\lambda\beta_1 \sum_{i=\frac{n}{2}}^{n} \max(0, -\sum_{j=1}^{\frac{n}{2}} \alpha_j^* k({\mathbf{z}^{\mathbf{SVM}}}_j, \text{sg}[{\mathbf{z}^{\mathbf{L}}}_i]) - \rho^*)}^{\text{Gradient flow only through the } {\mathbf{z}^{\mathbf{SVM}}}_i} \\ &\underbrace{+\lambda\beta_2 \sum_{i=\frac{n}{2}}^{n} \max(0, -\sum_{j=1}^{\frac{n}{2}} \text{sg}[\alpha_j^*] k(\text{sg}[{\mathbf{z}^{\mathbf{SVM}}}_j], {\mathbf{z}^{\mathbf{L}}}_i) - \text{sg}[\rho^*])}_{\text{Gradient flow only through the } {\mathbf{z}^{\mathbf{L}}}_i} \end{aligned} \tag{7}$$

This formulation of L_{OgAE} allows separating the influence of the ${\mathbf{z}^{\mathbf{SVM}}}_i$ and the ${\mathbf{z}^{\mathbf{L}}}_i$. We argue that the term weighted by β_1, which gradient flows through the ${\mathbf{z}^{\mathbf{SVM}}}_i$, influences the frontier of the SVM, as it moves samples in directions such that it includes the misclassified ${\mathbf{z}^{\mathbf{L}}}_i$ in the frontier: we call this term the *expander*. The term weighted by β_2, which gradient flows through the ${\mathbf{z}^{\mathbf{L}}}_i$, will influence the misclassified ${\mathbf{z}^{\mathbf{L}}}_i$, as it moves the samples in directions such that they enter the boundary drawn by the ${\mathbf{z}^{\mathbf{SVM}}}_i$: we call this term the *compactor*.

C. Algorithm and implementation details

The whole training procedure can be performed in two different manners. The first one consists in training first the

auto-encoder with guidance from the OCSVM loss term, based on batches of normal samples $\mathbf{x}$, followed by a final OCSVM-training on all encoded normal data $\mathbf{z}$. While effectively being in two parts, the final OCSVM training is fairly quick and computationally inexpensive. The second manner is to use the already computed OCSVMs of the last M iterations and perform a mean of the M decisions functions. This second technique provides a one-step training of the model. In the reminder of the manuscript, we use the first technique.

The whole procedure is summarized in the algorithm presented in the supplementary material S-A and technical details given in S-B. Input batches can be of any type, including non-image if the the autoencoder is adapted. In this paper, we focus on anomaly detection in images and consider either $\mathbf{x}$ as a whole image (Experiment 1), or as an image patch (Experiment 2). When considering 2D image patches, at inference, as in [49] and [50], the central pixel of each patch is associated to a latent representation and then an anomaly score. In this paper, the anomaly score is computed as the signed distance of the latent representation to the estimated OCSVM support. Then a whole 2D anomaly map can be obtained by moving this patch in increments of 1 in all directions across the entire image and calculating the score of the central pixel for each position. A 3D score map can obtained by concatenating the 2D anomaly maps (see first row of Figure 3 for examples of anomaly score maps superposed with MRIs).

IV. Experiment 1: Digit distinction under corruptions

We propose a first use-case experiment to evaluate the performance of the proposed model in a controlled setting against state-of-the-art models. The proposed task is to evaluate if the models can correctly classify handwritten digits of the normal class versus digits of other classes when presented with a wide variety of corruption noises.

A. Experimental setup and dataset

1) Corrupted MNIST database: MNIST-C [4] is a corrupted variant of the MNIST dataset, designed to evaluate model robustness under distribution shifts. It applies 15 different types of corruptions, such as noise, blur, and geometric transformations, to the original MNIST digits images of dimension 28x28. In this experiment, we train the networks on a "normal" digit class corresponding here to digit *3*, under specific corruptions, here *identity*, *motion blur* and *translate*, and then evaluate the networks ability to distinguish "normal" from anomalous digits, corresponding to digit *8*, under another distribution of corruptions, here *stripe*, *canny edges* and *brightness*, as exemplified on Figure 6 of the supplementary material. The training set is composed of 18393 images, concatenating 6131 handwritten *3* training images of each corruptions type, (*identity*, *motion blur* and *translate*). 90% of these images of the normal class are used for model training and 10% are used for early stopping. The validation set is composed of both 974 handwritten *8* and 1010 *3* images from the testing set of MNIST, corrupted with the testing corruptions (*stripe*, *canny edges* and *brightness* in Figure 6), for a total of 5952 images. The testing set is composed of both 5851 handwritten *8* and 6131 *3* images from the training set of MNIST, different from those of the training dataset and corrupted with the testing corruptions (*stripe*, *canny edges* and *brightness*), for a total of 35946 images. Using the original testing set for validation and the original training set for testing allows to give the testing set the most samples and thus the most statistical power. Note that the validation set and the testing set have no samples in common. The digit *8* was chosen to resemble digit *3* and thus correspond to a challenging classification task.

2) Compared methods: To evaluate our proposed method, we benchmark it against a set of commonly used approaches in UAD that align with the two main paradigms discussed in Section II: reconstruction-based methods (Section II-A) and support estimation-based methods (Section II-B), both using autoencoders for representation learning. First, we include standard Autoencoder (AE), Variational Autoencoder (VAE) and Siamese Autoencoder (SAE) models, assessing their anomaly detection performance based on reconstruction error. These models are widely used as baseline approaches in anomaly detection, as discussed in [5] and [11]. Additionally, we evaluate their combination (non-coupled) with a OCSVM trained on the learned latent space, following prior works [24], [60], [49]. Second, we include comparison with the standard coupled methods described in (Section II-B2), namely Deep SVDD (both *hard* and *soft* versions) [3], Deep Structure Preservation SVDD (DSPSVDD) [25], and Deep VAE-SVDD (DVAESVDD) [26]

For each method, we benchmark a set of their corresponding hyperparameters, and choose the best performing hyperparameter on the validation set. Performances are then reported for the testing set. The benchmarked hyperparameters are detailed in supplementary materials S-C.

3) Proposed task: We propose to perform a classification task at the image level. The different models take as input the whole image of dimension 28x28. They either directly output an anomaly score or output an anomaly score map of the same dimension as the input image (for reconstruction methods), that is then averaged to obtain a final score.

4) Metrics and statistical testing: We evaluate anomaly detection performance at the image level using standard metrics including the areas under the ROC curve (*AUROC*) and *AUROC30* which focuses on the low false-positive rate regime ($\leq$30%), as well as the Area Under the Precision-Recall Curve (*AUPR*), which is particularly relevant for highly imbalanced datasets. We perform statistical testing among the compared models, by generating 1000 bootstrap samples by sampling the testing set with replacement, then computing the evaluation metrics for each model on each bootstrap sample, and finally identifying the best-performing model based on the mean metric values. We then perform a paired bootstrap test, computing p-values as the fraction of bootstrap samples where a competing model outperforms the best model. To account for multiple comparisons, we apply Bonferroni correction, adjusting the significance threshold accordingly.

TABLE I: PERFORMANCE OF STUDIED MODELS ON DISCRIMINATING *3* VS *8* UNDER CORRUPTION. BEST MODEL IN BOLD. MODELS WITH NO STATISTICALLY SIGNIFICANT DIFFERENCE (P-VALUE < 0.01 AFTER PAIRED BOOTSTRAP TEST WITH BONFERRONI CORRECTION) ARE UNDERLINED.

3 vs 8	*AUROC*	*AUPR*	*AUROC30*
AE *recons. error*	0.56	0.66	0.66
AE + *ocsvm*	0.54	0.65	0.67
VAE *recons. error*	0.54	0.65	0.66
VAE + *ocsvm*	0.52	0.63	0.65
SAE *recons. error*	0.55	0.65	0.67
SAE + *ocsvm*	0.53	0.64	0.66
O*g*AE [ours]	<u>0.59</u>	**0.70**	**0.71**
h-DSVDD [3]	0.51	0.62	0.65
s-DSVDD [3]	0.52	0.63	0.66
DSPSVDD [25]	0.51	0.62	0.65
DVAESVDD [26]	**0.59**	0.67	0.65

B. Results and discussion

Table I presents the performance metrics obtained by all benchmarked models when distinguishing *3* from *8* under noise distribution shifts. We observe that performance of representation models coupled with OCSVM (AE + *ocsvm*, VAE + *ocsvm* and SAE + *ocsvm*) are on par with their reconstruction-based counterparts (AE *recons. error*, VAE *recons. error* and SAE *recons. error*). Our proposed model, O*g*AE, achieves better performances than any other model on all metrics (except when being on par with DVAESVDD for *AUROC*). Overall, the basic methods (AE-based) remain competitive, consistently performing within 5 points of the best model for every metric.

When comparing all coupled models (O*g*AE, h-DSVDD, s-DSVDD, DSPSVDD and DVAESVDD), we find that O*g*AE and DVAESVDD outperform their competitors and that on this non-trivial task, some coupled models are outperformed by basic baselines (AE *recons. error*), aligning with previous findings [17]. For deep SVDD, the results consistently show that the hard-margin variant of Deep SVDD (h-SVDD) outperforms or at least matches the performance of the soft-margin version (s-SVDD). This aligns with the original paper results [3] and the literature, which suggests that the added complexity of the soft-margin approach does not translate into a performance gain [25], [26], [29], [27]. Additionally, we find that DVAESVDD consistently outperforms DSPSVDD. This could highlight the advantage of using the VAE for more compact latent space or adapting the center of the hypersphere at each batch. Both methods consistently outperform traditional Deep SVDD approaches, aligning with the findings in their original papers [26], [25].

A global analysis of the results suggests several global patterns. Models that leverage representation learning combined with explicit support estimation generally outperform or are on par with reconstruction-based methods. Coupled approaches, where representation learning and anomaly detection are jointly optimized, tend to yield better results than decoupled methods. Recent methods that build upon Deep SVDD frameworks demonstrate improved performance over earlier variants. Finally, our proposed method achieves superior results on this benchmark, surpassing existing state-of-the-art models. We found the best performing hyperparameters for our method to be $\lambda = 1^{-3}$ and $(\beta_1, \beta_2) = (1, 0)$ for the first half of epochs (*expander*) and $(\beta_1, \beta_2) = (0.5, 0.5)$ for the last half of epochs (balanced *expander* + *compactor*).

It is worth noting that the dataset corruptions introduced in our experiments can be interpreted as a form of domain shift (i.e. corruptions in the test set are not the same as those in the training set), further emphasizing the adaptability of the evaluated models in real-world scenarios.

V. EXPERIMENT 2: LESION DETECTION IN BRAIN MRI

In this experiment, we consider a clinically realistic scenario and evaluate the models capabilities to detect brain lesions or tumors in MRI scans. As described in Section II-C, this setup is more challenging than the one in experiment 1, as brain MRI images typically contain more complex structures and noise, making anomaly detection more difficult. We consider two different applications that cover the spectrum of detection tasks mostly encountered in this domain. The first one consists in detecting and segmenting brain tumors, which are anomalies of large size but heterogeneous patterns, while the second one focuses on the detection on small and more subtle vascular lesions.

A. Experimental setup and dataset

We consider four 3D MRI T1 image databases described below, two databases of normal control subjects (V-A2), one for training and one for testing and two pathological databases for testing only (V-A1), the first one comprising exams of brain tumor patients from the public BraTS dataset [61] and the second one containing exams of patients with punctuate vascular lesions from the public WMH dataset [62]. To make the setup more challenging, we only use the T1 modality for each of these databases, unlike most studies of the literature which also combine the FLAIR images, where the lesions appear as hyperintense. Also note that the training and the testing databases are totally separated, introducing domain shift, contrary to some studies which train on the healthy slices of a database and test on the pathological ones. All databases undergo the same preprocessing procedure, described in the supplementary material S-D, to obtain 3D volumes of size $186 \times 218 \times 135$ with 1mm^3 voxel size.

1) Patient databases: The **BraTS2020 dataset** [61] from the 2020 brain tumor segmentation challenge consists of multi-modal brain MRI scans (T1, T1ce, T2 and FLAIR) of glioblastoma patients along with their 4-label tumor segmentation masks: background, enhancing tumor (ET), tumor core (TC) and whole tumor (WT). This dataset has also been recently employed to benchmark UAD models as described in section II-C. In this analysis, we randomly selected 60 out of the 369 available T1 MRI training images of the patient cohort with a mean age of 61.2 ± 11.8 years. Examples of 2D transverse slices extracted from the 3D volume are presented in left column of Figure 3.

The **White Matter Hyperintensities (WMH) dataset** originating from the WMH Segmentation Challenge [62] consists of 60 T1w and FLAIR MRI scans from patients exhibiting

small vessel diseases and acquired from three different hospitals (20 per hospital), along with expert-annotated segmentation masks. This dataset has been recently employed for the evaluation of unsupervised anomaly detection [36], [37], [63], [58], [50], using separate normative datasets for training and leveraging WMH data exclusively for evaluation. Examples of transverse 2D T1 MRI transverse slices extracted from the 3D volume are presented in left columns of Figure 4. The patient cohort has a mean age of 70.1 ± 9.3 years and exhibits a wide range of lesion volumes (0.78 cm³ to 195.15 cm³), making it particularly challenging due to inter-subject variability and scanner differences.

2) Control databases: The first normative database used for the training and validation of the different UAD models is the semi-public **CERMEP control dataset** [64] which comprises 75 healthy controls' MRI scans. The subjects in this control group have an average age of 38 ± 11.5 years, which is younger compared to the WMH and BraTS patient cohorts. Note that this age gap may introduce a domain shift between the train and test groups potentially impacting detection performance, notably as the process of normal-brain aging results in a slight brain shrinkage. The second control dataset used for testing is a subset of the openly available **IXI dataset**, which comprises nearly 600 MRI scans from healthy subjects. For this study, we selected 60 IXI controls for testing, age-matched to the WMH dataset (70.1 ± 9.3 years), to mitigate potential age-related bias in the model's classification performance. We considered that the age-distribution of the IXI dataset was close enough to that of the BraTS dataset (61.2 ± 11.8 years) to avoid extracting a second subset of the IXI dataset that would perfectly match the age distribution of the BraTS dataset.

3) Compared Methods: Based on the synthetic review (section II-C) of the state-of-the-art UAD approaches that have been evaluated on MRI of large and heterogeneous brain lesions, such as BraTS or more subtle lesions such as MSLUB, WMH or ATLAS (considering the small lesions only), we chose to pick one among the best performing models of each category. Considering support/density estimation methods, we selected the model combining a global siamese autoencoder and localized per-voxel OCSVM models [49], referred to as SAE + *localized* OCSVM in the following. This patch-based model used an autoencoder architecture similar to the one used in experiment 1 and detailed in the supplementary material S-E. We include the UNET-based autoencoder (AE) architecture proposed by Baur et al. [36] as well as the restoration model VQ-VAE + Transformer proposed by Pinaya et al. [37] which combines a quantized autoencoder (VQ-VAE) with an autoregressive transformer in the latent space. Both AE and VQ-VAE+Transformer models process full 2D slices which are concatenated to obtain the 3D anomaly score map. Finally, we evaluate performance of the cDDPM+MHD diffusion model proposed by Behrendt et al [44] which was shown to outperform DAE or anoDDPM and other masked-based models (e.g. autoDDPM [46]) for the detection of BraTS or small stroke lesions of the ATLAS dataset.

For a fair comparison, we implement each method using the hyperparameters provided in their respective publications. Hyperparameters for our proposed method are taken to be the best performing for experiment 1 to reduce the computational burden and provide a fairer comparison to the other methods (no hyperparameter optimization performed).

For the patch-based models, (SAE + *localized* OCSVM [49] and our proposed model O*g*AE with *localized* OCSVM), we set the 2D patch size to 15×15, as done in our previous work [49], [50] and so as to approximately match the size of the images on experiment 1. The 2D anomaly score maps are reconstructed as described in section III-C. The 3D score map is obtained by concatenating the 2D anomaly maps. For our proposed method (O*g*AE), we use batch of co-localized patches, hence the name O*g*AE with *localized* OCSVM.

4) Proposed tasks: We consider two different tasks in this experiment, a classification task at the image-level (3D), following the evaluation protocol of the first experiment in section IV-A3, as well as a localization (segmentation) task at the voxel-level. Both evaluations are derived from the single anomaly score map output by each model.

For the **classification** task, to obtain a single anomaly score per patient from their anomaly score map, we tested different aggregation methods (2% percentile, mean, median, with or without ventricle removal) and found that it had little impact on the overall results. In the end, we used the 2nd percentile threshold of the anomaly scores (meaning 2% of scores fall below this value) while excluding the ventricles[2]. For the **localization** task, we directly used the anomaly score maps and compared them voxel-wise to the ground-truth masks of the lesions, to obtain localization metrics.

Training of the models was done on 80% of the CERMEP control dataset, while the remaining 20% was used for early stopping during training. Testing was performed on the control IXI-age-matched subset, on the pathological WMH database and on the BraTS database for the classification task. For the localization task, testing was done on WMH and BraTS databases only (IXI controls indeed contain no lesions).

5) Metrics and statistical testing: We use the same evaluation metrics as in experiment 1 (*AUROC*, *AUROC30*, and *AUPR*) as detailed in section IV-A4, both for the classification task (distinguishing controls from patients) and the localization task (identifying lesions within patient images). Unlike the first experiment with balanced classes, this setup introduces imbalance in the localization task, where lesion voxels are rare. *AUPR* is thus critical, as it better reflects performance under imbalanced training. For the **classification** task, we perform statistical testing following the paired bootstrap test with Bonferroni correction described in section IV-A4 with 1000 bootstrap samples. For the **localization** task, we compute each metric per patient, thus introducing natural variability across samples (patients). We employ a Kruskal-Wallis test to detect overall differences among models, followed by Dunn's test for pairwise comparisons with Bonferroni correction.

TABLE II: CLASSIFICATION PERFORMANCE. BEST MODEL IN BOLD. MODELS WITH NO STATISTICALLY SIGNIFICANT DIFFERENCE (P-VALUE <0.01 AFTER PAIRED BOOTSTRAP TEST WITH BONFERRONI CORRECTION) ARE UNDERLINED.

Methods	Classification IXI vs BraTS		
	AUROC	*AUROC 30*	*AUPR* (0.5)
AE/UNet [36]	**1.00**	**1.00**	**1.00**
VQ-VAE + Transformer [37]	1.00	1.00	1.00
cDDPM + MHD [44]	1.00	1.00	1.00
SAE + *localized* OCSVM [49]	0.96	0.94	0.96
O*g*AE with *localized* OCSVM [ours]	0.87	0.83	0.89

Methods	Classification IXI vs WMH		
	AUROC	*AUROC 30*	*AUPR* (0.5)
AE/UNet [36]	0.98	0.97	0.98
VQ-VAE + Transformer [37]	0.95	0.93	0.96
cDDPM + MHD [44]	0.87	0.85	0.90
SAE + *localized* OCSVM [49]	**0.99**	**0.98**	**0.99**
O*g*AE with *localized* OCSVM [ours]	0.91	0.91	0.94

TABLE III: LOCALIZATION PERFORMANCE. BEST MODEL IN BOLD. MODELS WITH NO STATISTICALLY SIGNIFICANT DIFFERENCE (P-VALUE <0.01 AFTER KRUSKAL-WALLIS AND DUNN WITH BONFERRONI CORRECTION) ARE UNDERLINED.

Localization BraTS		
AUROC	*AUROC 30*	*AUPR* (0.007)
0.63	0.50	0.042
0.75	0.64	0.088
0.86	**0.79**	**0.247**
0.55	0.59	0.081
0.60	0.64	0.113

Localization WMH		
AUROC	*AUROC 30*	*AUPR* (0.007)
0.39	0.42	0.005
0.53	0.51	0.008
0.84	**0.73**	0.055
0.63	0.60	0.018
0.61	0.71	**0.066**

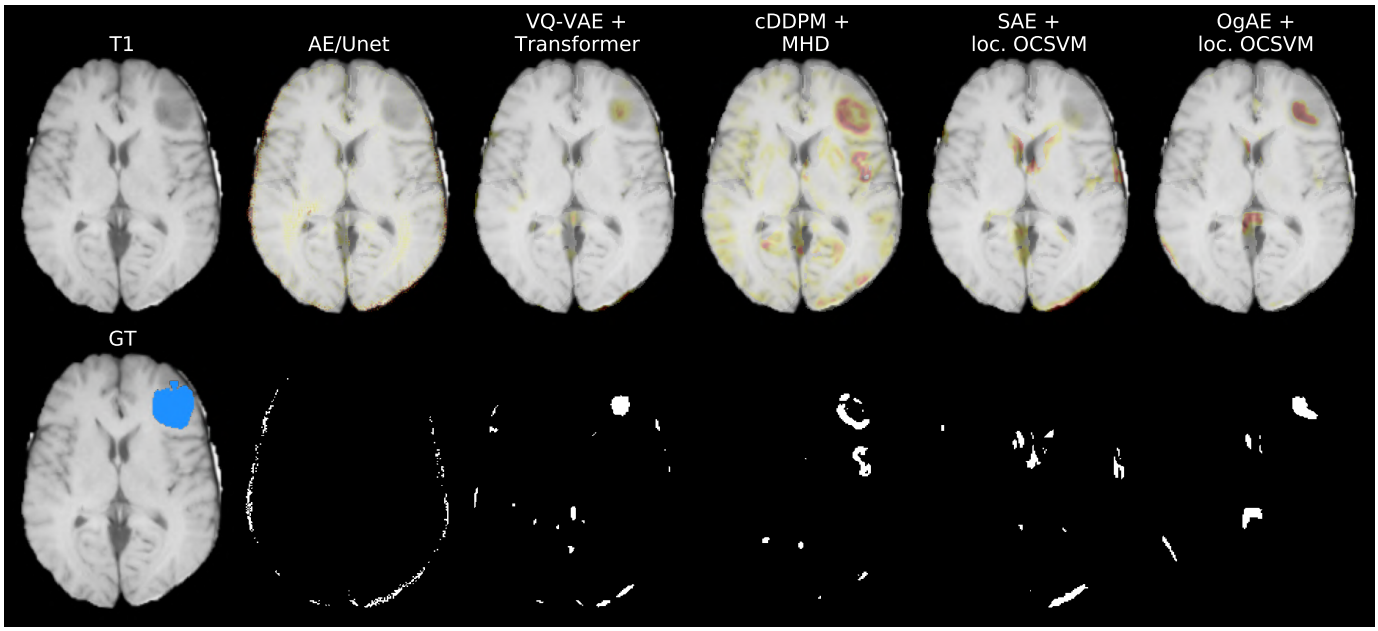

Fig. 3: Visualization of a central slice from the T1-weighted brain MRI of a BraTS patient (16). The ground truth (GT) is overlaid, with light blue indicating pathological lesions. Anomaly score maps from the studied methods are superimposed, with redder colors corresponding to higher anomaly scores. At the bottom, the anomaly map is thresholded at the 2% quantile.

B. Results and discussion

Classification and localization performance on the BraTS dataset are presented in table II and III and in Figure 3 (for the AE/UNet method, the dynamic range of the image had to be enhanced to [5%, 95%] quantile for enhanced visibility.). The three models based on reconstruction error (AE/UNet, VQ-VAE + Transformer and cDDPM + MHD) achieve perfect classification score. Our O*g*AE with *localized* OCSVM model achieves reasonable performance but significantly lower than the three best performing models. Performance of SAE+*localized* OCSVM is lower than the best performing models. Regarding localization performance, the diffusion cDDPM + MHD model is shown to significantly outperform all other models. Our proposed O*g*AE model ranks second regarding the *AUPR* and *AUROC30* metrics. As expected, and confirming results reported in the literature, simple AE/UNet has the lowest performance. These quantitative results are confirmed by the example score maps reported on Figure 3, highlighting the clear localization of the tumor in the anomaly score maps generated by the cDDPM + MHD (column 4) albeit balanced by a false detection, unlike our proposed O*g*AE model (column 6) where the false detections are of smaller size and located close to the brain boundaries or the ventricles. Localization of these false detections may suggest that our models are likely to be slightly impacted by registration errors. These false detections, especially when close to the brain boundaries could be erased by simple post-processing.

Classification and localization performance on the WMH dataset is presented in Tables II and III and Figure 4 (Dynamic range of AE/UNet had to be enhanced to [5%, 95%] quantile

[2]This exclusion was first motivated because ventricles tend to exhibit high anomaly scores due to age-related differences between the control and patient databases.

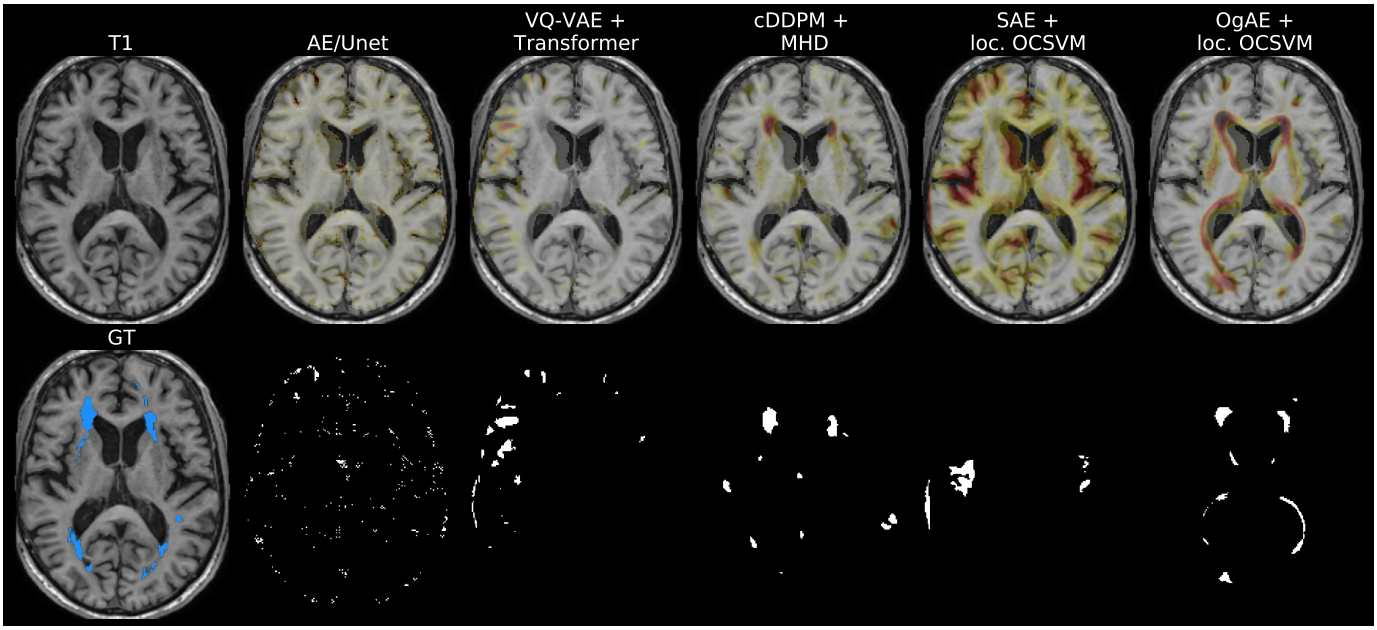

Fig. 4: Visualization of a central slice from the T1-weighted brain MRI of a WMH patient (AM126). The ground truth (GT) is overlaid, with light blue indicating pathological lesions. Anomaly score maps from the studied methods are superimposed, with redder colors corresponding to higher anomaly scores. At the bottom, the anomaly map is thresholded at the 2% quantile.

also.). All models achieve high accuracy, yet lower than performance achieved on the BraTS dataset, which can be explained by the higher difficulty of the task. We indeed emphasize that the detection and localization task on the WMH T1 MRI dataset is very challenging as exemplified on Figure 4 where the lesions highlighted in blue (bottom line of column 1) are small and very subtle, barely invisible on the original T1 MRI (first line of column 1). Regarding the classification task, although SAE+*localized* OCSVM emerges as the best-performing method, its advantage over other approaches is not statistically significant, except regarding cDDPM + MHD which is shown to significantly underperform compared to all other models. It is important to recall that the IXI and WMH test databases are age-matched, meaning that models should not be able to distinguish images based solely on age-related degenerative changes. This ensures that any detected anomalies are not confounded by age effects. Note that the rank ordering of the different methods for this classification task is different from that achieved for classification of the BraTS dataset, with a significantly improved performance of the two support estimation models and a decrease of performance for the reconstruction-based models. Comparison of the different models based on their localization performance is slightly different, with cDDPM + MHD and our proposed model O*g*AE with *localized* OCSVM ranking first, with cDDPM + MHD producing the best *AUROC* metric (0.84) significantly higher than that achieved by our model (0.61), while our model achieves the hightest *AUPR* value of 0.066, yet not statistically better than the *AUPR* achieved by cDDPM + MHD (0.055). Note that for this localization task, the baseline *AUPR* (random classifier) is 0.007. AE/UNet and VQ-VAE+Transformer are not capable of localizing correctly the lesions, as their performance are at chance level or below. This result is surprising, as these models perform well at the subject level, meaning they are capable of distinguishing between control and patient, but not by directly identifying the lesions' localizations. This could suggest that these models may have found other discriminant anomalies than those annotated by the clinicians, or other confounding features enabling discriminating the IXI from the WMH subjects. Example visualization on Figure 4 is in par with the quantitative results reported in table III. Simple AE/UNet and VQ-VAE+Transformer do not detect the subtle WMH lesions, while our O*g*AE with *localized* OCSVM model produces masks best matching the ground truth lesion masks outlined in blue. cDDPM + MHD also exhibits good sensitivity to the two upper lesions but fails at detecting the thinner bottom lesions, while VQ-VAE + Transformer and SAE+*localized* OCSVM estimate higher anomaly scores in cortical regions depicting a slight age-related shrinkage. This observation may be explained by the higher sensitivity of these models to the age-shift between the normative and the WMH distributions discussed in section V-A2.

In this study, we used the T1 MRI modality, where lesions are challenging to detect, unlike all previous studies performed on this database which also included MR FLAIR images where WMH lesions appear as hyperintense [63], [37], [36], [50], [58]. A broader trend emerges where models initially designed to detect hyperintense lesions struggle with this task. SAE+*localized* OCSVM, originally developed for epileptogenic lesions detection (which even experts struggle to see [65]), performs better in this context. Overall, our proposed method outperforms state-of-the-art methods or perform on par on this difficult task, particularly for identifying small lesions.

VI. General Discussion and conclusion

In this work, we introduced a novel method for UAD that addresses limitations of existing approaches: most state-of-the-art methods rely either on reconstruction-based models, which have to compromise sensitivity to the anomalies and specificity impaired by the imperfect reconstruction of normal tissue, or on decoupled support/density estimation architectures where feature learning and anomaly scoring are optimized separately resulting in misaligned feature spaces. Recent attempts to couple these processes often rely on surrogate objectives, linear kernel formulations, or approximations that compromise flexibility and robustness. To overcome these challenges, we proposed a coupled framework in which the representation learning process is explicitly guided by an analytically solvable OCSVM loss that steers the encoder toward producing latent features aligned with the OCSVM decision boundary, thereby directly optimizing the feature space for anomaly detection. By enforcing this alignment during training, the encoder is encouraged to focus on features that are genuinely relevant for modeling the normative distribution, reducing overfitting to irrelevant patterns.

We evaluated our approach on two tasks: digit distinction under corruption, and lesion detection in brain MRI. In the first task, our proposed O*g*AE model outperformed both classical and state-of-the-art UAD methods, on a task evaluating robustness to domain shifts across diverse corruptions. Early experiments where the training and testing corruptions would be the same was found too easy to discriminate the different UAD methods in this analysis (e.g. simple AE achieved $AUROC > 0.83$). Also, the setting where each model must distinguish between uncorrupted and corrupted digits is also fairly easy, with basic methods such as autoencoder reconstruction error reaching near perfect accuracy [5]. Additionally some corruptions were found to naturally project to the same latent space locations, thereby making the density/support estimation trivial and the reconstructions naturally erase the corruptions, thus, to provide a challenging setup, the corruptions used have been selected such that when training a basic autoencoder, they would each be separated in its latent space, which we verified using UMAP [66].

In the medical imaging task, we considered two challenging detection and localization tasks in T1 MRI, one on large and heterogeneous cancer lesions from the BraTS datasets and the second one regarding the highly difficult task of detecting small and subtle WMH lesions (barely invisible to naked eye). On the BraTS dataset, comparison with state-of-the art UAD models of different categories showed that our proposed O*g*AE model outperformed or matched existing models, with the exception of cDDPM + MHD. For the detection of small subtle WMH lesions, our proposed model outperformed the compared models, as confirmed by the improved *AUPR* metric. These results demonstrate, for the first time, the efficiency of coupled support estimation models compared to the reconstruction-based error models for a highly challenging clinical detection task. Additionally, at fixed hardware, our proposed method was found to train $\sim 2\times$ faster than cDDPM + MHD and infer $\sim 100\times$ faster.

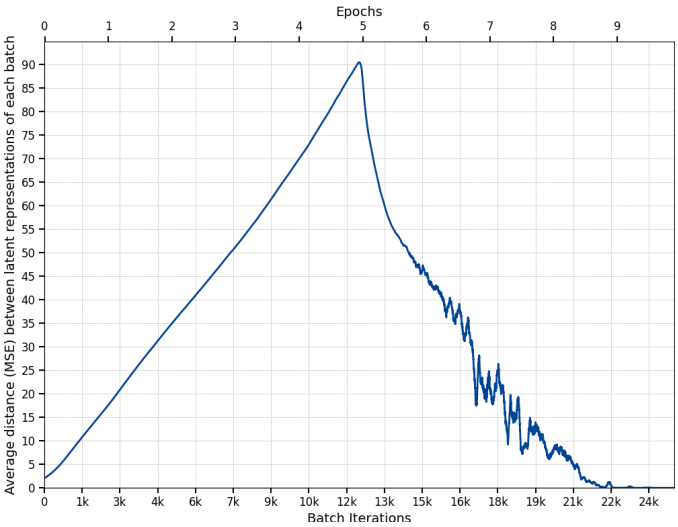

Fig. 5: Average MSE between latent representations during the training of the O*g*AE model for experiment 2. For the first 5 epochs the expander term is used, followed by the compactor term.

Additional computational resource analysis are performed in Supplementary material S-F.

A key contribution of our work is the OCSVM-guided representation learning, which addresses the limitations of existing coupled approaches: it avoids the pitfalls of traditional deep SVDD approaches, which often suffer from hypersphere collapse, by ensuring that the learned representations maintain sufficient variance while still being well-clustered within the normal class. In deep SVDD, soft-margin methods explicitly model the dual-space projection through a neural network, reducing expressivity, also, the widely used hard-margin variant focuses on compacting points around a predefined center without a notion of radius. Our approach, in contrast, does not rely on a neural network projection, preserving the full expressivity of the original OCSVM formulation. Furthermore, unlike methods that arbitrarily steer all points toward a center, our model allows them to remain in place if they lie within the estimated boundary, ensuring a sufficient level of variance in the learned representation. Additionally, we think computing the loss on a holdout portion of each batch can enhance generalization.

We show, on Figure 5, an example of training with the *expander* term (equation 7 $\beta_1 = 1, \beta_2 = 0$) for the first 5 epochs followed by the *compactor* term ($\beta_1 = 0$, $\beta_2 = 1$) for 5 other epochs. We study the average pairwise MSE between the latent representations, which is an indicator of their spread. We see that during the expanding phase the spread of the latent representation is growing and that in the compaction phase it is decreasing, proving what we intuited. The best performing strategy (evaluated on experiment 1) was found to be expander term first followed by expander + compactor with the same weight, aligning with the intuition that increasing the representation's variety at first benefits learning, but ultimately, the boundary size must be controlled and fixed. We believe further research is needed to explore optimal training strategies.

For the medical image experiment, we did not employ any

post-processing for our approach, unlike other works (e.g. AE/UNet in [36], VQ-VAE + Transformer in [37], cDDPM + MHD in [44]), suggesting that further refinement could improve performances, particularly in the localization task. Additionally, transitioning to 3D representations for medical images could enhance the model's spatial awareness. Previous research [50] suggests that patch size has minimal impact on performance, reinforcing the generalizability of our approach. Given that the SAE+*localized* OCSVM method [49] was effective for epilepsy detection, we should evaluate the potential of our proposed O*g*AE on epilepsy datasets as well, e.g. [67].

Several avenues for future research remain open. While our study focused on autoencoders, the OCSVM-guided framework could be applied to other feature extraction methods (e.g. transformers). Additionally, since we have focused our study on support estimation models (see section II-B), exploring density estimation techniques, which have proven competitive in anomaly detection, could provide further insights. Our method was designed for UAD (training only on normal samples), but in a semi-supervised setting, it could be extended by incorporating anomalous samples to refine the decision boundary: instead of only enforcing that normal samples remain inside the estimated boundary, anomalous samples could be explicitly pushed outside (or the frontier compacted such the sample remain outside). Also, an SVDD-guided variant could be implemented and evaluated, despite being similar when using the RBF kernel.

REFERENCES

[1] I. Lagogiannis, F. Meissen, G. Kaissis, and D. Rueckert, "Unsupervised pathology detection: a deep dive into the state of the art," *IEEE transactions on medical imaging*, vol. 43, no. 1, pp. 241–252, 2023.

[2] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the Support of a High-Dimensional Distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.

[3] L. Ruff *et al.*, "Deep one-class classification," in *International conference on machine learning*. PMLR, 2018, pp. 4393–4402.

[4] N. Mu and J. Gilmer, "Mnist-c: A robustness benchmark for computer vision," *arXiv preprint arXiv:1906.02337*, 2019.

[5] L. Ruff *et al.*, "A Unifying Review of Deep and Shallow Anomaly Detection," *Proceedings of the IEEE*, vol. 109, pp. 756–795, May 2021.

[6] M. A. Kramer, "Autoassociative neural networks," *Computers & chemical engineering*, vol. 16, no. 4, pp. 313–328, 1992.

[7] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proceedings of the MLSDA, Machine Learning for Sensory Data Analysis*. ACM, 2014, pp. 4–11.

[8] L. Beggel, M. Pfeiffer, and B. Bischl, "Robust anomaly detection in images using adversarial autoencoders," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany*. Springer, 2020, pp. 206–222.

[9] N. Pinon, "Unsupervised anomaly detection in neuroimaging: Contributions to representation learning and density support estimation in the latent space," Ph.D. dissertation, INSA Lyon, 2024, chapter III, Sec 1.

[10] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9592–9600.

[11] 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, 2021.

[12] D. Zimmerer, J. Petersen, and K. Maier-Hein, "High-and low-level image component decomposition using vaes for improved reconstruction and anomaly detection," *arXiv preprint arXiv:1911.12161*, 2019.

[13] L. Wang, D. Zhang, J. Guo, and Y. Han, "Image Anomaly Detection Using Normal Data Only by Latent Space Resampling," *Applied Sciences*, vol. 10, no. 23, p. 8660, Dec. 2020.

[14] X. Chen, N. Mishra, M. Rohaninejad, and P. Abbeel, "Pixelsnail: An improved autoregressive generative model," in *International conference on machine learning*. PMLR, 2018, pp. 864–872.

[15] H. He *et al.*, "A diffusion-based framework for multi-class anomaly detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 8, 2024, pp. 8472–8480.

[16] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," in *Advances in Neural Information Processing Systems*, 2018, pp. 9758–9769.

[17] Y. Cai, W. Zhang, H. Chen, and K.-T. Cheng, "Medianomaly: A comparative study of anomaly detection in medical images," *Medical Image Analysis*, vol. 102, p. 103500, 2025.

[18] A. Kascenas, R. Young, B. S. Jensen, N. Pugeault, and A. Q. O'Neil, "Anomaly detection via context and local feature matching," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[19] J. Tan, B. Hou, J. Batten, H. Qiu, B. Kainz *et al.*, "Detecting outliers with foreign patch interpolation," *Machine Learning for Biomedical Imaging*, vol. 1, no. April 2022 issue, pp. 1–27, 2022.

[20] Y. Zhang *et al.*, "Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features," *Proceedings of the AAAI*, vol. 35, no. 9, pp. 10 777–10 785, 2021.

[21] D. M. Tax and R. P. Duin, "Support Vector Data Description," *Machine Learning*, vol. 54, no. 1, pp. 45–66, Jan. 2004.

[22] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "Padim: a patch distribution modeling framework for anomaly detection and localization," in *International Conference on Pattern Recognition*. Springer, 2021.

[23] P. Perera and V. M. Patel, "Learning deep features for one-class classification," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5450–5463, 2019.

[24] S. Mabu, S. Hirata, and T. Kuremoto, "Anomaly detection using convolutional adversarial autoencoder and one-class svm for landslide area detection from synthetic aperture radar images." *J. Robotics Netw. Artif. Life*, vol. 8, no. 2, pp. 139–144, 2021.

[25] Z. Zhang and X. Deng, "Anomaly detection using improved deep SVDD model with data structure preservation," *Pattern Recognition Letters*, vol. 148, pp. 1–6, Aug. 2021.

[26] Y. Zhou, X. Liang, W. Zhang, L. Zhang, and X. Song, "Vae-based deep svdd for anomaly detection," *Neurocomputing*, vol. 453, 2021.

[27] H. Hojjati and N. Armanfard, "Dasvdd: Deep autoencoding support vector data descriptor for anomaly detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 8, pp. 3739–3750, 2023.

[28] H.-J. Xing and P.-P. Zhang, "Contrastive deep support vector data description," *Pattern Recognition*, vol. 143, p. 109820, 2023.

[29] J. Yi and S. Yoon, "Patch svdd: Patch-level svdd for anomaly detection and segmentation," in *Computer Vision-ACCV 2020: 15th Asian Conference on Computer Vision, Kyoton*. Springer, 2021, pp. 375–390.

[30] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[31] B. Zong *et al.*, "Deep autoencoding gaussian mixture model fore unsupervised anomaly detection," in *International Conference on Learning Representations (ICLR)*, 2018, p. 19.

[32] P. Perera, R. Nallapati, and B. Xiang, "Ocgan: One-class novelty detection using gans with constrained latent representations," in *Proceedings of the IEEE/CVF*, 2019, pp. 2898–2906.

[33] F. Behrendt, D. Bhattacharya, L. Maack, J. Krüger, R. Opfer, and A. Schlaefer, "A review of deep learning-based unsupervised anomaly detection in brain mri," *Medical Image Analysis*, vol. 112, p. 104076, 2026.

[34] Ž. Lesjak *et al.*, "A novel public mr image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus," *Neuroinformatics*, vol. 16, pp. 51–63, 2018.

[35] O. Commowick *et al.*, "Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure," *Scientific reports*, vol. 8, no. 1, p. 13650, 2018.

[36] C. Baur, B. Wiestler, M. Muehlau, C. Zimmer, N. Navab, and S. Albarqouni, "Modeling Healthy Anatomy with Artificial Intelligence for Unsupervised Anomaly Detection in Brain MRI," *Radiology: Artificial Intelligence*, vol. 3, no. 3, p. e190169, May 2021.

[37] W. H. Pinaya *et al.*, "Unsupervised brain imaging 3D anomaly detection and segmentation with transformers," *Medical Image Analysis*, vol. 79, p. 102475, Jul. 2022.

[38] V. M. Muñoz-Ramírez, V. Kmetzsch, F. Forbes, and M. Dojat, "Deep Learning Models to Study the Early Stages of Parkinson's Disease," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. Iowa City, IA, USA: IEEE, Apr. 2020, pp. 1534–1537.

[39] D. Zimmerer, F. Isensee, J. Petersen, S. Kohl, and K. Maier-Hein, "Unsupervised anomaly localization using variational auto-encoders," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2019 Shenzhen*. Springer, 2019, pp. 289–297.

[40] Y. Zhao, Q. Ding, and X. Zhang, "Ae-flow: Autoencoders with normalizing flows for medical images anomaly detection," in *The Eleventh International Conference on Learning Representations*, 2022.

[41] Z. Liang *et al.*, "Itermask3d: Unsupervised anomaly detection and segmentation with test-time iterative mask refinement in 3d brain mri," *Medical Image Analysis*, p. 103763, 2025.

[42] A. Kascenas *et al.*, "The role of noise in denoising models for anomaly detection in medical images," *Medical image analysis*, vol. 90, p. 102963, 2023.

[43] J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks, "AnoDDPM: Anomaly Detection with Denoising Diffusion Probabilistic Models using Simplex Noise," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 649–655.

[44] F. Behrendt *et al.*, "Leveraging the mahalanobis distance to enhance unsupervised brain mri anomaly detection," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 394–404.

[45] F. Behrendt *et al.*, "Guided reconstruction with conditioned diffusion models for unsupervised anomaly detection in brain mris," *Computers in Biology and Medicine*, vol. 186, p. 109660, 2025.

[46] C. I. Bercea, M. Neumayr, D. Rueckert, and J. A. Schnabel, "Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models," in *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*, 2023.

[47] C. I. Bercea, B. Wiestler, D. Rueckert, and J. A. Schnabel, "Diffusion models with implicit guidance for medical anomaly detection," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, M. G. Linguraru *et al.*, Eds. Cham: Springer Nature Switzerland, 2024, pp. 211–220.

[48] F. Beizaee, G. Lodygensky, C. Desrosiers, and J. Dolz, "Mad-ad: Masked diffusion for unsupervised brain anomaly detection," in *Information Processing in Medical Imaging*, I. Oguz, S. Zhang, and D. N. Metaxas, Eds. Cham: Springer Nature Switzerland, 2026, pp. 139–153.

[49] Z. Alaverdyan, J. Jung, R. Bouet, and C. Lartizien, "Regularized siamese neural network for unsupervised outlier detection on brain multiparametric magnetic resonance imaging: Application to epilepsy lesion screening," *Medical Image Analysis*, vol. 60, p. 101618, 2020.

[50] N. Pinon, R. Trombetta, and C. Lartizien, "One-Class SVM on siamese neural network latent space for Unsupervised Anomaly Detection on brain MRI White Matter Hyperintensities," in *MIDL 2023, International Conference on Medical Imaging with Deep Learning*. PMLR, 2023.

[51] N. Pinon, G. Oudoumanessah, R. Trombetta, M. Dojat, F. Forbes, and C. Lartizien, "Brain subtle anomaly detection based on auto-encoders latent space analysis : application to de novo parkinson patients," in *ISBI 2023 - IEEE 20th International Symposium on Biomedical Imaging*. Cartegena de Indias, Colombia: IEEE, Feb. 2023.

[52] M. El Azami, A. Hammers, J. Jung, N. Costes, R. Bouet, and C. Lartizien, "Detection of lesions underlying intractable epilepsy on t1-weighted mri as an outlier detection problem," *PloS one*, vol. 11, 2016.

[53] C. Bowles *et al.*, "Brain lesion segmentation through image synthesis and outlier detection," *NeuroImage: Clinical*, vol. 16, pp. 643–658, 2017.

[54] B. H. Menze *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[55] A. Kascenas, N. Pugeault, and A. Q. O'Neil, "Denoising autoencoders for unsupervised anomaly detection in brain mri," in *International Conference on Medical Imaging with Deep Learning*. PMLR, 2022, pp. 653–664.

[56] F. Beizaee, S. Hajimiri, I. Ben Ayed, G. Lodygensky, C. Desrosiers, and J. Dolz, " Reflect: Rectified Flows for Efficient Brain Anomaly Correction Transport ," in *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, vol. LNCS 15963. Springer Nature Switzerland, September 2025.

[57] 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*, 2021.

[58] F. Meissen, G. Kaissis, and D. Rueckert, "Challenging current semi-supervised anomaly segmentation methods for brain mri," in *International MICCAI brainlesion workshop*. Springer, 2021, pp. 63–74.

[59] N. Pinon, "Unsupervised anomaly detection in neuroimaging : Contributions to representation learning and density support estimation in the latent space," Theses, INSA de Lyon, Apr. 2024.

[60] N. Pinon, R. Trombetta, and C. Lartizien, "Détection d'anomalies dans l'espace image ou l'espace latent d'auto-encodeurs par patch pour l'analyse d'images industrielles," in *GRETSI 2023, XXIXème Colloque Francophone de Traitement du Signal et des Images*, 2023.

[61] S. Bakas *et al.*, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific Data*, vol. 4, 2017.

[62] H. J. Kuijf *et al.*, "Standardized Assessment of Automatic Segmentation of White Matter Hyperintensities and Results of the WMH Segmentation Challenge," *IEEE Transactions on Medical Imaging*, vol. 38, no. 11, pp. 2556–2568, Nov. 2019.

[63] W. H. Pinaya *et al.*, "Fast unsupervised brain anomaly detection and segmentation with diffusion models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 705–714.

[64] I. Mérida *et al.*, "CERMEP-IDB-MRXFDG: a database of 37 normal adult human brain [18F]FDG PET, T1 and FLAIR MRI, and CT images available for research," *EJNMMI Research*, vol. 11, p. 91, Dec. 2021.

[65] T. Wehner *et al.*, "Factors influencing the detection of treatable epileptogenic lesions on mri. a randomized prospective study," *Neurological research and practice*, vol. 3, pp. 1–11, 2021.

[66] L. McInnes, J. Healy, N. Saul, and L. Großberger, "Umap: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[67] F. Schuch *et al.*, "An open presurgery mri dataset of people with epilepsy and focal cortical dysplasia type ii," *Scientific Data*, vol. 10, 2023.

[68] A. Agrawal, B. Amos, S. Barratt, S. Boyd, S. Diamond, and Z. Kolter, "Differentiable convex optimization layers," in *Advances in Neural Information Processing Systems*, 2019.

[69] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, Apr. 2011.

[70] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd, "Osqp: An operator splitting solver for quadratic programs," *Mathematical Programming Computation*, vol. 12, no. 4, pp. 637–672, 2020.

SUPPLEMENTARY MATERIAL

A. Algorithm

We present two ways of implementing our proposed O*g*AE model, one with a final OCSVM training and another with a storage of the last M OCSVMs. Both have proven to be similar in terms of performances, lightweight and fast, with as few as $M = 10$.

Autoencoder training with OCSVM-guidance

Input: Normal samples $(\mathbf{x}_i)_{1\leq i\leq N}$
Output: Trained encoder E ▷ Decoder D is discarded

for each epoch **do**
  for every batch of samples $(\mathbf{x}_i)_{1\leq i\leq b}$ **do** ▷ Batch size b
    Compute latent representations of samples :
    $(\mathbf{z}_i)_{1\leq i\leq b} = E[(\mathbf{x}_i)_{1\leq i\leq b}]$
    Split in two the $\mathbf{z}_i$ to obtain ${\mathbf{z}^{\text{SVM}}}_i$ and ${\mathbf{z}^{\text{L}}}_i$
    Solve the OCSVM problem for the ${\mathbf{z}^{\text{SVM}}}_i$ to obtain: $(\alpha_j^*)_{1\leq j\leq \frac{b}{2}}$ and ρ^*
    Compute the reconstructions of latent representations:
    $(\hat{\mathbf{x}}_i)_{1\leq i\leq b} = D[(\mathbf{z}_i)_{1\leq i\leq b}]$
    Compute the loss (7) and apply a gradient step to E and D

    if iteration $\in$ M last iterations **then**
      Save $(\alpha_j^*)_{1\leq j\leq \frac{b}{2}}$ and ρ^*
    end if
  end for
end for

OCSVM final training

Input: Normal samples $(\mathbf{x}_i)_{1\leq i\leq N}$ and trained encoder E
Output: Decision function f of OCSVM

Compute latent representations of samples:
$(\mathbf{z}_i)_{1\leq N} = \mathbf{D}[(\mathbf{x}_i^h))_{1\leq N}]$
Solve the OCSVM problem for the $(\mathbf{z}_i)_{1\leq N}$ to obtain the parameters of the final decision function

---

In the case of final OCSVM training, the final decision (and the encoder) is readily available for inference. In the case of storage of the last M OCSVMs, the mean of the M decisions functions is performed at inference.

B. Technical details for the OCSVM-guidance model

This section outlines the technical implementation of the OCSVM-guidance model, particularly the gradient computation through the dual solution, the numerical stabilization techniques, and the kernel matrix reformulation.

When computing the expander term in equation 7, we have to differentiate through $\boldsymbol{\alpha}^*$ and ρ^*, thus through a convex optimization problem (the dual problem): to do this we use [68]. . For solver-related manner, this dual problem has to be written in a way that it is linear in parameters, not quadratic. We thus utilize the fact that $\mathbf{K}$ is positive semi-definite (because it is a gram matrix), to express it as: $\mathbf{K} = {\mathbf{K}^{\frac{1}{2}}}^T \mathbf{K}^{\frac{1}{2}}$. Where $K_{ij} = k(\mathbf{z}_i, \mathbf{z}_j)$. As recommended in [69], because

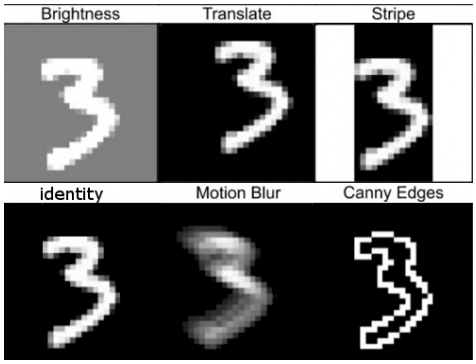

Fig. 6: Corruptions of the MNIST dataset (MNIST-C [4]) used throughout this article, on the digit 3.

$\frac{1}{\nu_j n}$ can get very small as n increase, this only leaves a tight bound for the constraint $0 \leq \alpha_{ji} \leq \frac{1}{\nu_j n}$. Thus, for numerical stability reasons, we solve a scaled problem of variable $\tilde{\alpha}_{ji} = n\nu_j \alpha_{ji}$. As also recommended in [69] for numerical stability, to compute ρ, we average the ρ value obtained for every support vector. Finally, also for numerical stability, we computed $\mathbf{K}$ as $\mathbf{K} + 1e^{-8}\mathbf{I}$. We used the OSQP solver [70].

C. Benchmarked hyperparameters for experiment 1

- Weight coefficient for KL divergence (VAE and DVAESVDD): $\beta_{\text{KL}} \in \{1, 10^{-1}, 10^{-2}\}$
- Weight coefficient for cosine similarity (SAE): $\alpha \in \{1, 10^{-1}, 10^{-2}\}$
- λ (O*g*AE): $\lambda \in \{1, 10^{-1}, 10^{-2}, 10^{-3}\}$
- γ (DSPSVDD, see article [25]): $\gamma \in \{10^{-1}, 1, 10\}$ (balance coefficient)
- α (DVAESVDD, see article [26]): $\alpha \in \{10^{-1}, 1, 10\}$ (balance coefficient)
- (β_1, β_2), i.e. expander/compactor strategy (O*g*AE) : $(\beta_1, \beta_2) \in \{(1, 0), (0, 1), (0.5, 0.5)\}$ or $(0, 1)$ for the first half epoch followed by $(0.5, 0.5)$ or $(0, 1)$ for the first half epoch followed by $(1, 0)$
- ν (OCSVM): $\nu \in \{0.01, 0.03, 0.1, 0.3, 0.5\}$
- γ_{RBF} (OCSVM): $\gamma_{\text{RBF}} \in \{10^{-1}, 10^{-2}, 10^{-3}\}$
- Scaling of latent variables (every method using OCSVM): True or False

The same autoencoder and training procedure are used for every method, to ensure fair comparison. The architecture is the one used for the MNIST experiment in DVAESVDD [26], it is detailed in the supplementary material S-E1.

D. Brain MRI registration and preprocessing pipeline for experiment 2

The brain MRI T1 preprocessing applied in this paper is based on a pipeline implemented in SPM12 and fully described in [49]. This pipeline includes a critical registration step that enables precise voxel-wise comparisons across subjects by aligning all images to a standardized anatomical space. Spatial normalization was performed using the unified segmentation algorithm (UniSeg) which includes segmentation of grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF),

correction for magnetic field inhomogeneities and spatial normalization to the standard brain template of the Montreal Neurological Institute (MNI). In this work, we used the default parameters for normalization and a voxel size of $1 \times 1 \times 1$ mm. The cerebellum and brain stem were excluded from the spatially normalized images. The masking image in the reference MNI space was derived from the Hammersmith maximum probability atlas. On top of that, each image was intensity-normalized with: $X_{\text{norm}} = \frac{X - \min(X)}{\max(X) - \min(X)}$. To account for the specificity of the BraTS data which are skull-stripped, we considered two versions of the CERMEP and IXI normative datasets, one where images contain signal from the skull for performance analysis on the WMH dataset and one where images are skull-stripped to match the preprocessing applied on the BraTS dataset. Also, as the downloadable BraTS images are corrected from magnetic field inhomogeneities and skull-stripped, we thus applied an affine registration step to the MNI template to complete the preprocessing.

E. Autoencoder architectures

1) Experiment 1: The autoencoder architecture for all models of experiment 1 is the one used for the MNIST experiment in DVAESVDD [26]. It consists of a convolutional encoder and a symmetric decoder. The encoder comprises two convolutional layers (5×5 kernels, 4 and 8 filters), each followed by batch normalization, LeakyReLU activation, and 2×2 max pooling. The latent representation is obtained via a fully connected layer of dimension 32 (meaning reduction factor of 24.5). The decoder mirrors the encoder, employing a dense layer to reshape the latent space, followed by two transposed convolutional layers (5×5 kernels, 8 and 4 filters) interleaved with batch normalization, LeakyReLU activation, and 2×2 upsampling. A final transposed convolution (5×5, 1 filter) with a sigmoid activation reconstructs the input. The model is trained with mean squared error as the reconstruction loss, optimized with Adam (learning rate: 1e-3), with a batch size of 100, for 20 epochs.

2) Experiment 2: The encoder consists of four convolutional layers: a 5×5 layer with 3 filters, followed by three successive 3×3 layers with 4, 12, and 16 filters, respectively. Each convolutional layer is paired with batch normalization and GELU activation. The decoder mirrors this structure precisely. It begins with three 3×3 transposed convolutional layers with 12, 4, and 3 filters, each followed by batch normalization and GELU activation, and concludes with a 5×5 transposed convolution and a sigmoid activation. For training, we optimize the model using mean squared error (MSE) with the Adam optimizer (learning rate: 1e-3), trained for 10 epochs with a batch size of 100.

F. Hardware setup and computation performances analysis

The majority of the computations were run on a local computer equiped with a GeForce GTX 1660 SUPER (6Gb VRAM), along with 16 Gb of RAM and and AMD Ryzen 5 3600 6-Core Processor. Occasionally when launching large batches of experiments, a HPC equiped with a NVIDIA Tesla V100 (32Gb VRAM) was used.

For Experiment 1, for all methods, among training of the autoencoder model, training of the final ocsvm (if there was one) and inference, the training of the autoencoder was always the longest of the three. OCSVM training time (when applicable) was below 1.37s for every method (only using CPU) and inference time was between 2.27s and 8.12s depending on the models (longest being AE and quickest DVAESVDD). Training of the OgAE took the longest time with 465.49s, with other methods being between 38.77s and 64.47s. Mean GPU power during training was between 37.0W for SAE and 61.1W for OgAE. Mean CPU and RAM usage were very similar (14% difference at max for RAM and 18% at max for CPU) between each methods. We believe that these computation cost values while measured on Experiment 1, still holds for experiment 2 (for the patch-based methods), where the patches used are smaller (15x15 in Experiment 2 against 28x28 in Experiment 1).

For Experiment 2, as stated in section VI, our proposed method was found to train $\sim 2\times$ faster than cDDPM + MHD (7.5h VS 13h) and infer $\sim 100\times$ faster (3.20min = 0.05h VS 5h). Training time of the AE/UNet was 4.9h and VQ-VAE + Transformer was 18.6h (10h for VQ-VAE and 8.6h for Transformer). As our proposed method is patch-based, we expect that its computational cost will only grow linearly as a function of the image size, contrary to methods that take as input full images.