Artificial Intelligence In Medicine 143 (2023) 102606

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Research paper

Self-supervised out-of-distribution detection in wireless capsule endoscopy images

Arnau Quindós, Pablo Laiz, Jordi Vitrià, Santi Seguí *

Departament de Matemàtiques i Informàtica, Universitat de Barcelona (UB), Barcelona, Spain

ARTICLE INFO

Keywords:
Out-of-distribution
Anomaly detection
Deep learning
Capsule endoscopy

ABSTRACT

While deep learning has displayed excellent performance in a broad spectrum of application areas, neural networks still struggle to recognize what they have not seen, i.e., out-of-distribution (OOD) inputs. In the medical field, building robust models that are able to detect OOD images is highly critical, as these rare images could show diseases or anomalies that should be detected. In this study, we use wireless capsule endoscopy (WCE) images to present a novel patch-based self-supervised approach comprising three stages. First, we train a triplet network to learn vector representations of WCE image patches. Second, we cluster the patch embeddings to group patches in terms of visual similarity. Third, we use the cluster assignments as pseudolabels to train a patch classifier and use the Out-of-Distribution Detector for Neural Networks (ODIN) for OOD detection. The system has been tested on the Kvasir-capsule, a publicly released WCE dataset. Empirical results show an OOD detection improvement compared to baseline methods. Our method can detect unseen pathologies and anomalies such as lymphangiectasia, foreign bodies and blood with $AU\,ROC > 0.6$. This work presents an effective solution for OOD detection models without needing labeled images.

1. Introduction

Wireless capsule endoscopy (WCE) is an endoscopy technique that is an alternative to the standard procedure originally developed by Iddan et al. [1]. This method presents a variety of advantages versus standard endoscopy due to being far less invasive, not requiring sedation, and risking fewer potential complications. WCE makes use of a small pill-shaped capsule with a camera inside, rather than the traditional long, thin, flexible tube with a camera at one of its ends. This capsule can be easily swallowed, upon which the camera records hours of intestinal video that a medical team can later view to diagnose any gastrointestinal condition.

Nevertheless, WCE videos can contain thousands of images per patient and must be screened by medical specialists. This is a time-consuming and complex process. Its repetitive nature might lead to missing pathologies or other important elements [2]. For this reason, artificial intelligence offers a clear opportunity to support this task [3–5].

The application of AI techniques has been thoroughly investigated for the detection of abnormal or suspicious images in WCE. Several works have been presented for the identification or segmentation of pathological conditions such as bleeding [6–8], polyps or tumors [9–14], angiectasia [15], ulcers [16–18], motility disorders [19], as well as methods for multipathology detection [20–26]. Deep learning currently represents the state-of-the-art for most of these problems and has demonstrated promising results. Nevertheless, independent of the performance on the task for which these models were designed, the ability to detect unseen out-of-distribution (OOD) images is crucial, as such OOD images may correspond to other severe conditions. For example, a polyp is an abnormal growth of tissue that can evolve into cancer, and therefore, its detection can be highly beneficial. However, a system that accurately detects polyps but fails to identify advanced-stage tumors would not be desirable. Therefore, the development of reliable ODD detectors in addition to supervised detectors is necessary for adoption in clinical practice.

The nature of capsule endoscopy images is wide and heterogeneous, which challenges deep learning models to learn what is normal or in-distribution. Furthermore, some images are considered abnormal due to an anomaly in a small area of the image, despite the remaining image being completely normal. In these cases, an OOD detector will most likely classify those as in-distribution, as the anomaly cannot outweigh the in-distribution features of the image. Therefore, one of the goals of this work is to develop a detector that is able to identify small anomalies (see Fig. 1).

In this study, we introduce a self-supervised method derived from ODIN [28] based on patches. We first create a model able to generate

* Corresponding author.
E-mail address: santi.segui@ub.edu (S. Seguí).

https://doi.org/10.1016/j.artmed.2023.102606
Received 22 August 2022; Received in revised form 2 June 2023; Accepted 5 June 2023
Available online 15 June 2023

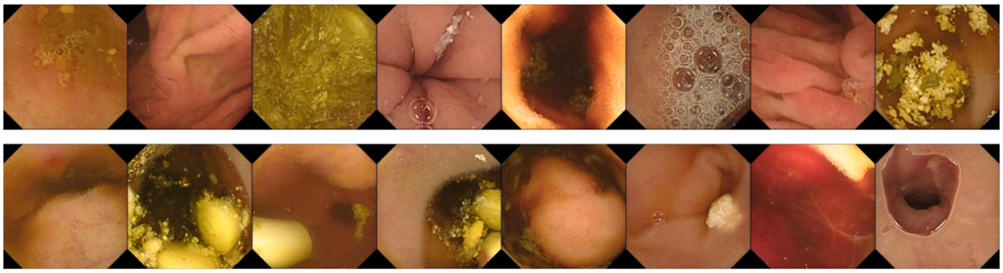


Fig. 1. Random WCE sample images that illustrate the diversity of the dataset [27] and the complexity of out-of-distribution detection. First row: normal frames. Second row: frames containing some pathology.

vector representations of fixed-size patches extracted from WCE images as a self-supervised task. These embeddings encode visual features from the patches and allow the creation of clusters of patches in terms of visual similarity. Finally, we train a classifier using cluster assignments as pseudolabels. Similar to its predecessor method ODIN, our OOD detector is based on the confidence of this patch classifier.

The remainder of this paper is structured as follows. First, we present an overview of the related work in the OOD field. Then, we describe the details of our methodology, followed by the experimental setup and results. Finally, we conclude the paper and provide directions for future work.

2. Related work

The OOD image detection problem in deep learning has been studied for many years using a variety of approaches ranging from conventional statistical techniques (such as density estimation) to generative models (such as autoencoders or GANs). In this study, we distinguish between supervised methods, which use some type of labeling in the training set, and self-supervised methods, which learn the necessary knowledge to perform the OOD problem without the need for specific labeling in the training set. Our method falls in the second category.

Supervised methods

A widely used baseline method for this problem is the maximum over softmax probabilities (MSP) [29]. This approach is based on a classifier trained over in-distribution data and works on the assumption that models make more confident predictions with in-distribution inputs than with OOD data. It conceptually depends on the outputs of a simple multiclass classifier and requires no further training. However, its performance has proven to be inferior to many later approaches. Thus, it is currently only used as a baseline method.

Since diverse and enormous datasets of images are available, Hendrycks et al. [30] proposed leveraging these data to improve OOD detectors against auxiliary datasets of outliers in a method called Outlier Exposure (OE). In this method, the classifier is trained to predict a uniform distribution over labels for outlier inputs, which enforces low confidence over these inputs. Thulasidasan et al. [31] proposed using an abstention class in the classification problem and assigning known outliers to this class. Further work showed that leveraging the labels of the known outliers instead of assigning all outliers to a single abstention class can further enhance the performance of the OOD detector, despite only representing a small subset of the type of outliers that we want to detect [32].

Another effective improvement to MSP is ODIN [28]. While still being a confidence-based approach, ODIN includes two fundamental novel techniques: temperature scaling and input perturbation. These techniques lead to better OOD detection, making it one of the best-performing state-of-the-art methods for the OOD problem. Nonetheless, ODIN relies on OOD data to tune the temperature and perturbation hyperparameters. In contrast, a generalized version of ODIN (GODIN [33]) does not require tuning with OOD data and mitigates this issue.

Other approaches focus on modeling the class-conditional distribution of pretrained CNN features with a Gaussian distribution and use the Mahalanobis distance in the predicted class distribution to detect OOD samples [34]. For example, DeepIF method [35] achieves better detection performance by modeling the distribution of CNN features with a nonparametric technique based on isolation forests.

Self-supervised methods

The concept of learning normality to then detect anomalies is evident in methods based on deep-generative models including autoencoders (AEs), variational autoencoders (VAEs) and generative adversarial networks (GANs). All of these methods learn features with high representation quality that can be used for density estimation methods [36] or reconstruction error methods [37]. These approaches rely on the assumption that reconstruction models trained on in-distribution images produces higher-quality outcomes with in-distribution inputs than with OOD inputs. Thus, images producing a high reconstruction error can be classified as OOD.

Other self-supervised approaches have tried to replicate classifier-based supervised methods without using labeled data, such as ensemble leave-out classifiers proposed by Vyas et al. [38]. This technique consists of randomly partitioning data in K subsets and creating K classifiers, each of which samples one of the K subsets without replacement as OOD data and samples the remaining subsets as in-distribution training data.

3. Methodology

The general concept of the proposed method in this paper is to use ODIN [28], which is considered one of the state-of-the-art approaches in OOD detection, to detect abnormal areas of WCE images. To focus on small regions of the image, we split the WCE images into fixed-size patches, which we consider our training and testing examples. Since ODIN is a classifier-based approach, labels are required to classify samples. Toward this application, we use a self-supervised feature extraction network to generate embeddings and then apply a clustering algorithm that assigns each sample one label that is later used to train the classifier. Importantly, our method does use any external labeling of the images or patches.

The pipeline of the method comprises three stages, which are henceforth described in detail.

Triplet-loss embeddings

The first stage of our method seeks to learn a vector representation for patches extracted from WCE images. To learn these embeddings, we use a triplet loss (TL) network, which allows us to perform self-supervised learning, as described in the next paragraph.

A TL architecture compares an anchor input with two other inputs: a positive input, which shares a property with the anchor, and a negative

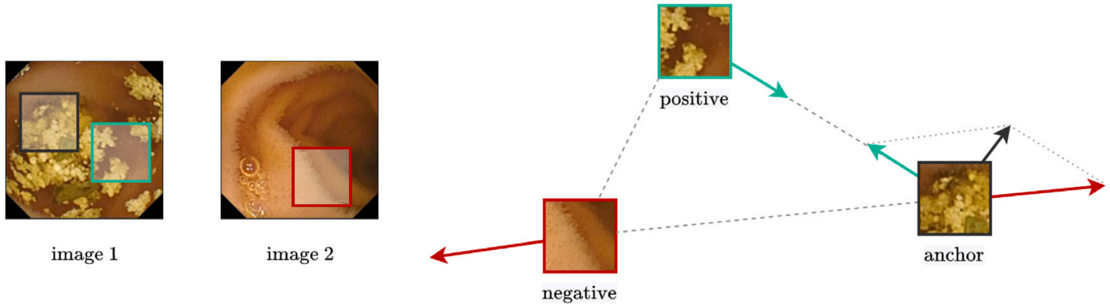


Fig. 2. Triplet loss applied to three patches that are transformed into three vectors. Anchor: patch from a given image. Positive: a different patch of the same image. Negative: a patch of a different image.

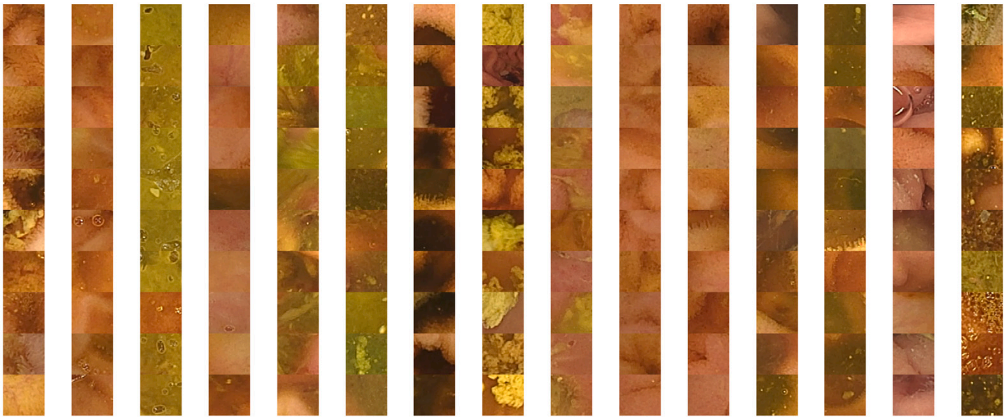


Fig. 3. Example of clusters produced; each column represents one cluster. Patches in the same cluster are more visually similar than patches in different clusters ($K = 15$, patch size 96 × 96). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

input, which does not share this property. In our case, inputs are fixed-size patches extracted from WCE images, and the shared property is that both subsets are extracted from the same image, whereas the negative patch is extracted from a different image, as illustrated in Fig. 2. TL aims at ensuring that the anchor image, x_a, is closer to all other images from the same class, x_p, than any image from a different class, x_n.

To achieve classification, the following loss function is used:

$$\mathcal{L}_{TL}(x_a, x_p, x_n) = \sum_{i=1}^{N} \max\left(\|f_i^a - f_i^p\| - \|f_i^a - f_i^n\| + \alpha, 0\right) \tag{1}$$

where f^k is the vector representation of x^k, N is the batch size, $\|\cdot\|$ is a norm and α is a margin parameter to enforce separation between classes.

Given two patches from the same image, this network generates embeddings that are closer together than two patches from different images. Since two patches from the same image will tend to be more visually similar than two from different images, these embeddings can be useful to cluster patches based on visual features.

Cluster pseudolabeling

As outlined above, we must label patches to train a patch classifier. Therefore, we use cluster predictions as pseudolabels to train our OOD-detector classifier.

Given the triplet-loss embeddings produced in the previous stage, we use the mini-batch K-means algorithm to create patch clusters based on visual similarity. Images in the same cluster tend to share visual features such as color, texture or shape. An example of such a clustering is shown in Fig. 3.

This clustering partitions the patch dataset, which is then used to train a K-class classifier.

Patch ODIN classifier

The third and final stage of our method is the patch-based ODIN, which is based on a K-class classifier trained with the aforementioned pseudolabels. This classifier also includes temperature scaling and input perturbation, as defined in the original ODIN paper [28]. A given image is partitioned into m patches and fed to the ODIN model, which outputs an anomaly score for each patch (see Fig. 4).

Formally, we let x be an input patch, $\tilde{x}$ be the perturbed version of this patch and $z = (z_1, \dots, z_K)$ be the output vector produced by the temperature-scaled K-softmax layer. We define the anomaly score, called the softmax score, in Eq. (2). Then, these scores are combined using a summary function to obtain a measure of the abnormality of the image as a whole. If this score surpasses a certain threshold, then the image is labeled as OOD.

$$\mathcal{S}(\tilde{x}; T) = 1 - \max_{i=1,\dots,k} softmax(\tilde{x}; T)_i = 1 - \max_{i=1,\dots,k} z_i \tag{2}$$

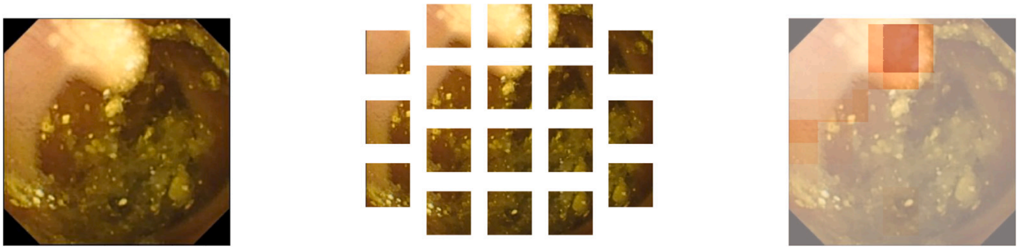
For each image x, we will extract m patches $x_1, \dots, x_m$. Given a perturbation magnitude ε, a temperature parameter T and a threshold δ, our OOD discriminator is defined as follows:

$$OOD(x; T, \varepsilon) = \begin{cases} 1, & \Psi\left(\mathcal{S}(\tilde{x}_1; T), \dots, \mathcal{S}(\tilde{x}_m; T)\right) \geq \delta \\ 0, & \Psi\left(\mathcal{S}(\tilde{x}_1; T), \dots, \mathcal{S}(\tilde{x}_m; T)\right) < \delta \end{cases} \tag{3}$$

where Ψ is a summary function applied to the softmax scores of the patches.

Given the softmax scores of the patches of an image $\vec{y} = (y_1, \dots, y_m)$, we define three summary functions in Table 1.

Each of these three strategies may perform differently depending on the nature of the anomalies to detect. For instance, *max*, which only uses the patch with the highest anomaly score, might work better for localized features but might also introduce more noise; *wavg*,



(a) WCE example image with lymphangiectasia

(b) Patches extracted from the image. Note some overlap between patches.

(c) Heatmap produced by ODIN. Red represents high anomaly score areas.

Fig. 4. Illustration of the patch-splitting process.

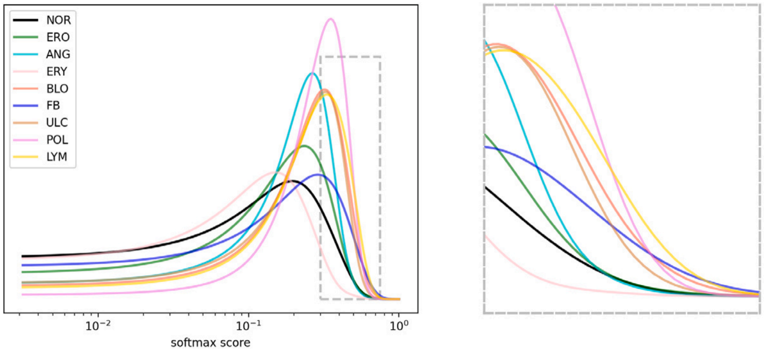


Fig. 5. Softmax score distributions produced by patch ODIN (*top5* summary) by pathology. Real distributions are used to fit Gaussian distributions, which are plotted above. The second window shows a zoomed view. Distributions are normalized for the sake of comparison.

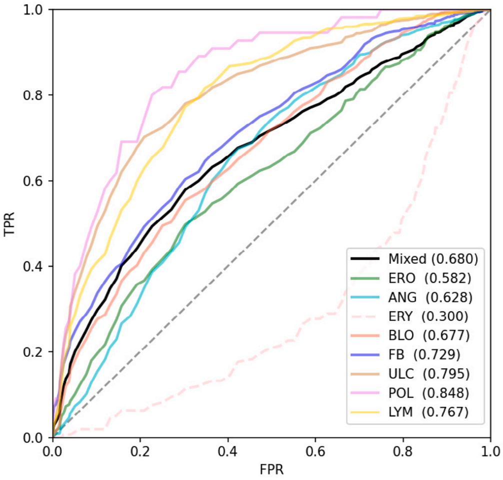


Fig. 6. ROC curve of patch ODIN (*top 5*) OOD detection by pathology. Mixed considers all Kvasir pathological frames as one single class.

which accounts for all the patches but gives more importance to the highest scores, might be better suited for global anomalies; *top-k* is an intermediate approach that might work like *max* but with less noise.

4. Experimental setup

4.1. Dataset

We evaluate and compare our proposed method with Kvasir-Capsule [27]. Kvasir-Capsule is a publicly released WCE dataset that contains 117 videos of gastrointestinal footage from different patients, 74 of which are unlabeled and 43 partially labeled. The labeled frames are comprised of 14 different classes, 5 of which refer to nonpathological categories: normal clean mucosa (NOR), ileocecal valve (IV), pylorus (PYL), reduced mucosal view (RED) and ampulla of vater (AV); and 9 that refer to pathological or abnormal categories: angiectasia (ANG), erythema (ERY), blood - fresh (BLO), blood - hematin (BLH),[1] erosion (ERO), foreign body (FB), ulcer (ULC), polyp (POL) and lymphangiectasia (LYM). For our OOD problem, we considered the 9 pathological categories as OOD, i.e., our detection target.

Table 1
Definition of the summary functions used in this paper.

max maximum-score patch	$\Psi_{\max}(y) = \max_{i=1,\ldots,m} y_i$
top-k average of top-k patches	$\Psi_{\text{top } k}(y) = \frac{1}{k}\sum_{y_i \in S_k}{}^{a}\ y_i$
wavg weighted average of all the patches	$\Psi_{\text{wavg}}(y) = \frac{1}{m}\sum_{y_i \in S_m}{}^{a}\ \lambda^i y_i$

[a]Where S_n is the subset of the first n softmax scores, sorted in descending order.

Different frames from the same video can be very similar. Thus, data partitions must be done by videos instead of frames. We randomly selected 64 out of the 74 unlabeled videos for the training of the triplet network, the K-Means clustering and the patch classifier. The 10 remaining unseen videos are used as an intermediate validation set to assess the quality of the resulting embeddings and clustering and the accuracy of the classifier. The 43 labeled videos are then used only for testing purposes, with normal classes considered in-distribution and pathological frames of OOD. We extract fixed-size patches of 96 × 96 pixels from a video frame resolution of 336 × 336 using a step size of 60 pixels between patches, while ensuring that there is overlap between patches and that all areas of the image are captured.

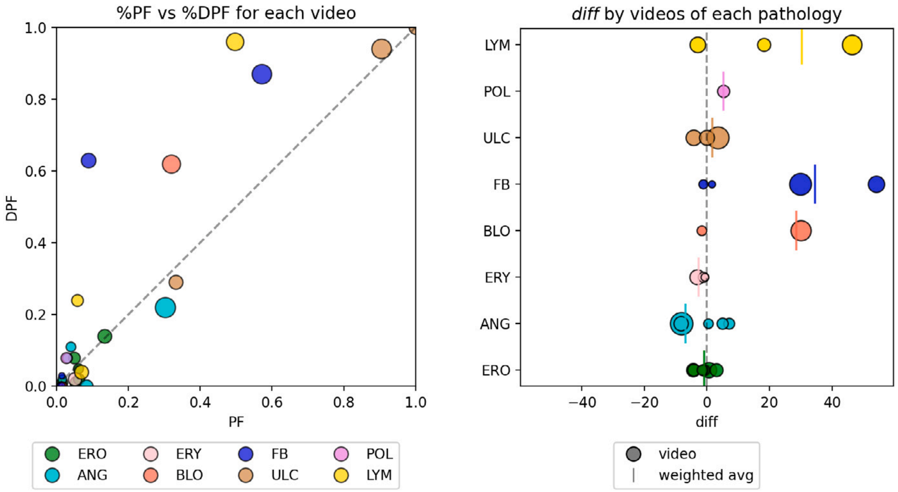
[1] BLH class is not considered for evaluation purposes due to the small number of frames available in this category, of which there are only 10.

Table 2
AUROC scores of OOD detection by pathology of the proposed Patch ODIN method. Comparison between three different patch sizes (PS).

Pathology	#samples	PS 64 × 64			PS 96 × 96			PS 128 × 128		
		max	*top-k*	*wavg*	*max*	*top-k*	*wavg*	*max*	*top-k*	*wavg*
ANG	866	0.576	0.629	0.636	0.572	0.628	0.635	0.592	0.591	0.595
BLO	446	0.646	0.708	0.701	0.613	0.677	0.678	0.644	0.616	0.601
ERO	507	0.579	0.593	0.598	0.562	0.582	0.587	0.538	0.557	0.564
ERY	159	0.412	0.398	0.407	0.322	0.301	0.300	0.457	0.468	0.479
FB	776	0.669	0.698	0.696	0.737	0.729	0.732	0.621	0.613	0.602
LYM	592	0.645	0.671	0.671	0.739	0.767	0.772	0.627	0.635	0.640
POL	55	0.531	0.417	0.435	0.776	0.848	0.845	0.580	0.600	0.600
ULC	854	0.615	0.618	0.626	0.725	0.795	0.804	0.534	0.521	0.518
Aggregated	4255	0.611	0.638	0.640	0.650	0.680	0.686	0.578	0.578	0.575

Table 3
AUROC scores of OOD detection by pathology, comparison between different methods. For each pathology, the best score is marked in bold.

Pathology	#samples	Patch ODIN *96 × 96, top k*	Patch VAE *96 × 96, top k*	ODIN	VAE	SelectiveNet
ANG	866	**0.628**	0.367	0.483	0.573	0.515
BLO	446	0.677	0.705	0.541	**0.791**	0.576
ERO	507	0.582	**0.622**	0.570	0.540	0.472
ERY	159	0.301	0.231	**0.560**	0.324	0.326
FB	776	**0.729**	0.623	0.642	0.679	0.632
LYM	592	**0.767**	0.671	0.752	0.745	0.738
POL	55	**0.848**	0.350	0.667	0.622	0.652
ULC	854	**0.795**	0.706	0.543	0.680	0.669
Aggregated	4255	**0.680**	0.577	0.578	0.642	0.572



(a) %PF vs. %DPF ($n = 100$)

(b) *diff* by pathology

Fig. 7. Results by video. Each circle represents a video, and the size is proportional to the number of frames contained in that video.

4.2. Method stages

1. **Triplet-loss embeddings** The triplet-loss model uses the EfficientNetB0 [39] architecture, followed by a global average pooling layer. Finally, a 1280-unit dense layer outputs the feature vectors.
2. **K-Means clustering** Due to the high volume of images, we used the mini-batch version of the K-Means algorithm. The value chosen for the number of clusters is $K = 20$ due to seemingly showing the most consistent clustering results.
3. **Patch ODIN classifier** We use a CNN classifier that comprises the first three blocks of the ResNet50v2 [40] architecture (ImageNet pretrained), followed by three fully connected layers integrating dropout and batch normalization layers. The top fully connected layer uses a temperature scaling with a fixed temperature parameter value $T = 1000$, as proposed in other previous work [33]. Because the number of labeled videos for each pathology is limited, cross-validating these parameters is very risky.

4.3. Baseline methods

We use baseline self-supervised methods trained with the same data to compare the performance of our approach with other previous work. The implementation of supervised approaches would require using additional data or labels and, more importantly, would make an unfair comparison. The following methods are used as baselines:

ODIN [28]. To incorporate a nonpatch-based ODIN as a baseline method, we trained a self-supervised model that uses temporal information of the frames to generate pseudolabels and train an ODIN classifier on the full image. Importantly, this adapted version is self-supervised, indicating that it has not been trained with ground-truth labels.

SelectiveNet [41]. The SelectiveNet architecture includes a rejection option for selective classification, which we use as the OOD score. We use the same self-supervised methodology to generate pseudolabels as described for ODIN.

VAE [37]. We use the same VAE architecture and train on the same in-distribution data as our approach. Then, we use the VAE anomaly score proposed in their work to determine if an image is an OOD.

Patch VAE. We use a patch-based method, but instead of ODIN, we introduce the VAE from the baseline method (trained on patches) to create an anomaly score for every patch. Then, for every image, the scores of the patches are combined using a summary function to obtain an anomaly score for the image. We use this model as an intermediate between the VAE and our method.

4.4. Evaluation metrics

- **Accuracy.** Measures the fraction of examples correctly classified.
- **True negative rate (TNR)** or **specificity**. Refers to the proportion of negative examples that were classified as negative: $TN/(TN+FP)$.
- **True positive rate (TPR)** or **sensitivity**. Refers to the fraction of positive examples that were classified as positive: $TP/(TP+FN)$.
- **AUROC**. Area under the ROC curve. Refers to the sensitivity-specificity tradeoff at various threshold settings. To determine AUROC, we compute the anomaly score of normal and OOD samples and measure sensitivity and specificity at TPR and FPR at different threshold configurations.
- **TPR at n% TNR**, abbr. TPRn. Refers to the TPR when the TNR is n%. TPR95 and TPR90 are used.
- **%PF**. Refers to the percentage of pathological frames among all the frames in a video.
- **%DPF(n)**. Refers to the percentage of pathology among the n frames with the highest outlier scores.
- **Difference between %PF and %DPF (*diff*)**. Given the high variance in %PF across different videos in the dataset, we use this to measure how well the model detects OOD frames for different pathological prevalence.

  $$diff = \%\text{PF} - \%\text{DPF}$$

In addition to these quantitative metrics, we also evaluate the system qualitatively by inspecting the results produced on a subset of images. We consider this a very important evaluation to understand the predictions obtained by each model, which allows us to understand which images the model considers abnormal and which others are classified confidently.

5. Results

In our first experiment, we seek to analyze anomaly score distributions produced by the patch ODIN[2] for each class. Toward this goal, we extract the softmax score of each image and fit a Gaussian distribution to each set of scores. Fig. 5 shows these normalized distributions, i.e., with balanced classes to better compare the degree of overlap of these distributions. We note that, as a result of the plain nature of ERY images in this dataset, this class produces even lower anomaly scores than the normal class. The normal class (NOR) produces the lowest anomaly scores and thus allows us to separate classes using these scores. However, the degree of separation varies for each pathology: some classes, such as ERO and BLO, have a large overlap with NOR, while others, such as LYM or ULC, have a significant separation. We compare each pathology versus the normal class to evaluate the potential of our OOD discriminator. The results for a patch size of 96 are shown using ROC curves in Fig. 6.

To further investigate the effect of the patch size, we repeated the experiments using additional patch sizes of 64 and 128. The results are presented in Table 2, which shows that the patch size of 96 yields the best results. Therefore, we adopt this size for the subsequent experiments.

The AUROC score by pathology and comparison with baseline methods are summarized in Table 3. The results show that, considering all pathological frames as one abnormal class, our method slightly improves performance over the baseline methods. Considering each pathology individually, we observe different results. Patch ODIN performs especially well with LYM and ULC and slightly outperforms the baseline methods with FB and ERO. However, our method does not improve BLO detection, which the VAE model does especially well. This is attributed to blood being the most global anomaly, such that splitting the data into patches does not contribute to better detection. Furthermore, we observed that the best summary strategy depends on pathology. Some of the pathologies are global, while others appear very localized; overall, *top5* and *wavg* seem to yield the best results.

In general, we observe that VAEs tend to assign higher anomaly scores to images that appear more complex in terms of texture, colors and shapes. For instance, we find that nonpathological bubble images are usually assigned high scores, while pathological plain images are not detected. This mainly occurs because complex features, despite being common in the training set, are harder to reconstruct for an autoencoder. Thus, reconstruction error is higher for complex images, which plays a large role in anomaly score.

The availability of images for certain pathologies is extremely limited (different images may be consecutive frames of the video that contain the same anomaly), which can lead to inaccurate results. For this reason, further qualitative and quantitative analysis is necessary to confirm the performance of the system.

Notably, to compute AUROC scores, we use normal and OOD frames extracted from different videos. In real-world situations, given a WCE video from a single patient, it is desirable to flag the most abnormal frames to detect any potential condition. To better measure the performance of the model in such a situation, we test our method in each video separately and measure how well the model detects pathological frames among the 'most abnormal' frames. For this test, we use the *diff* metric described in the previous section, with $n = 100$. This metric compares the percentage of pathological frames among the 100 frames with the highest outlier score (%DPF(100)) with respect to the percentage of pathological frames in the video (%PF).

The results of the video analysis are shown in Fig. 7. We observe that frames containing LYM, FB and BLO produce high anomaly scores and thus are detected among the most outlier frames. For remaining diseases, the average *diff* is close to 0, indicating that the model detects pathological frames (among the most abnormal) at the same rate as they are present throughout the video. Because labeled videos containing ULC are >90% pathological, *diff* may not be the best performance indicator for this class. Additionally, these results may not match with the AUROC scores as previously presented because this metric measures each video independently and focuses on the most abnormal tail end.

We also conducted a qualitative analysis using the outputs of the model on a subset of images. To do this, for each selected image, we examined the score of each patch, analyzed which patches produced the highest scores, and plotted the results in the form of a heatmap over the original image. This process is illustrated in Fig. 8, where the model performs well, and in Fig. 9, where the model fails to correctly identify anomalies. We examined both successful and unsuccessful examples to determine those types of anomalies our model is able to identify and those which it cannot. A general conclusion is that the model tends to detect more visually prominent anomalies more accurately, as was expected.

[2] For the patch ODIN, we fix the following parameters: $k = 20$, $\varepsilon = 5 \cdot 10^{-4}$, and $T = 1000$.

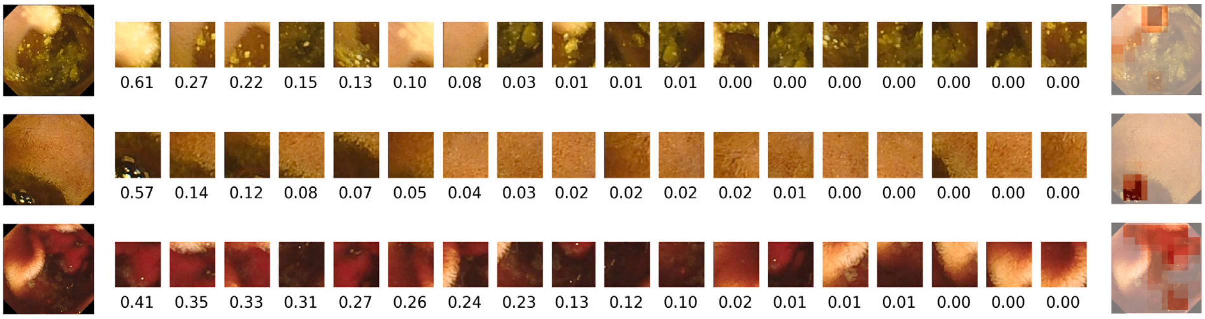


Fig. 8. Left: sample WCE images that contain LYM, FB and BLO, from top to bottom. Center: patches extracted from each image sorted in terms of softmax score. Right: Heatmaps produced using softmax scores; red areas represent high anomaly scores. In these examples, the model correctly identifies anomalies, and thus, patches containing anomalies produce high scores. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

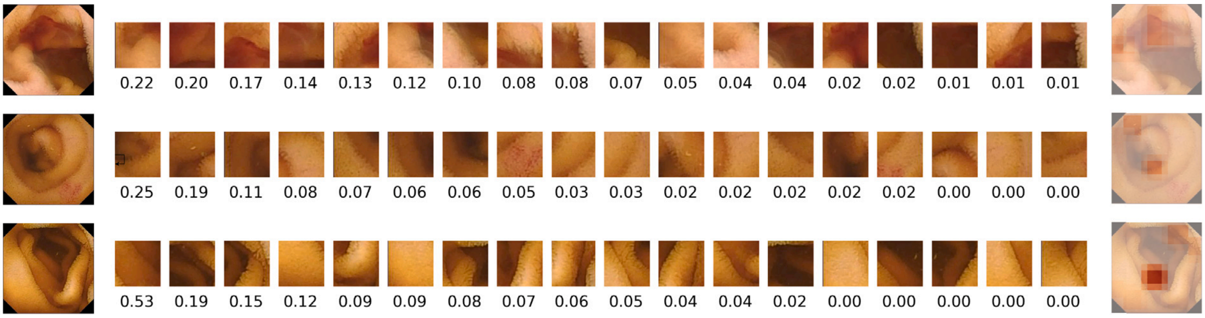


Fig. 9. Left: sample WCE images that contain BLO, ANG and no pathologies (NOR), from top to bottom. Center: patches extracted from each image sorted in terms of softmax score. Right: Heatmaps produced using softmax scores; red areas represent high anomaly scores. In these examples, the model is not able to correctly identify anomalies. In the first two cases, all the patches are assigned low scores, and thus, any abnormal area is detected. In the third case, the model incorrectly assigns a high score to a normal patch. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6. Conclusions

This study presents a method to improve OOD detection in WCE images with respect to other self-supervised approaches, such as VAEs or ODIN without patches. Both quantitative and qualitative results show that the system successfully detects pathologies including lymphangiectasia, foreign bodies and blood showing. Moreover, the patch-based nature of our methodology allows us to measure the abnormality of every region of the image.

While our method tends to effectively detect the most visually prominent anomalies, it is less sensitive to subtler anomalies such as erosion, angiectasia or erythema. These pathologies are visually quite similar to in-distribution WCE images. Therefore, detecting these types of anomalies using a model that has not been previously exposed to them is a great challenge. Moreover, the limited availability of data in medical fields reinforces the need for larger and more diverse datasets. Toward this goal, our future work may consider incorporating online learning techniques, where the model could dynamically adapt as a medical team flags unseen images to the system.

Overall, we intend that our method can provide an effective solution for OOD detection models without the need for labeled images. While this work has focused on WCE images, our methodology can also be applied to OOD detection in other computer vision applications.

Declaration of competing interest

None Declared.

Acknowledgments

This work has been partially funded by MINECO Grant RTI2018-095232-B-C21, 2021 SGR 01104 (Generalitat de Catalunya), Innovate UK project 104633, AI AWARD 02440 project from NIHR and NHSX and the FPU grant FPU17/02727 (Ministerio de Universidades, Spain).

References

[1] Iddan G, Meron G, Glukhovsky A, Swain P. Wireless capsule endoscopy. Nature 2000;405:417.

[2] Koulaouzidis A, Dabos K, Philipper M, Toth E, Keuchel M. How should we do colon capsule endoscopy reading: a practical guide. Ther Adv Gastrointest Endosc 2021;14:26317745211001983.

[3] Yang YJ. The future of capsule endoscopy: The role of artificial intelligence and other technical advancements. Clin Endosc 2020;53(4):387–94.

[4] Robertson AR, Segui S, Wenzek H, Koulaouzidis A. Artificial intelligence for the detection of polyps or cancer with colon capsule endoscopy. Ther Adv Gastrointest Endosc 2021;14:26317745211020277.

[5] Koulaouzidis A, Bjørsum T, Toth E. Real-life practice data on colon capsule endoscopy: We need them fast!. Endosc Int Open 2022;10(03):E230–1.

[6] Hajabdollahi M, Esfandiarpoor R, Soroushmehr SMR, Karimi N, Samavi S, Najarian K. Segmentation of bleeding regions in wireless capsule endoscopy images an approach for inside capsule video summarization. CoRR 2018. arXiv:1802.07788.

[7] Aoki T, Yamada A, Kato Y, Saito H, Tsuboi A, Nakada A, Niikura R, Fujishiro M, Oka S, Ishihara S, et al. Automatic detection of blood content in capsule endoscopy images based on a deep convolutional neural network. J Gastroenterol Hepatol 2020;35(7):1196–200.

[8] Saraiva MM, Ferreira JP, Cardoso H, Afonso J, Ribeiro T, Andrade P, Parente MP, Jorge RN, Macedo G. Artificial intelligence and colon capsule endoscopy: automatic detection of blood in colon capsule endoscopy using a convolutional neural network. Endosc Int Open 2021;9(08):E1264–8.

[9] Yuan Y, Li D, Meng MQ-H. Automatic polyp detection via a novel unified bottom-up and top-down saliency approach. IEEE J Biomed Health Inf 2017;22(4):1250–60.

[10] Laiz P, Vitrià J, Wenzek H, Malagelada C, Azpiroz F, Seguí S. WCE polyp detection with triplet based embeddings. Comput Med Imaging Graph 2020;86:101794.

[11] Yang J, Chang L, Li S, He X, Zhu T. WCE polyp detection based on novel feature descriptor with normalized variance locality-constrained linear coding. Int J Comput Assist Radiol Surg 2020;15(8):1291–302.

[12] Saito H, Aoki T, Aoyama K, Kato Y, Tsuboi A, Yamada A, Fujishiro M, Oka S, Ishihara S, Matsuda T, et al. Automatic detection and classification of protruding lesions in wireless capsule endoscopy images based on a deep convolutional neural network. Gastrointest Endosc 2020;92(1):144–51.

[13] Falin Z, Haihua L, Ning P. Gastrointestinal polyps and tumors detection based on multi-scale feature-fusion with WCE sequences. 2022, arXiv preprint arXiv:2204.01012.

[14] Gilabert P, Vitrià J, Laiz P, Malagelada C, Watson A, Wenzek H, Segui S. Artificial intelligence to improve polyp detection and screening time in colon capsule endoscopy. Front Med 2022;9.

[15] Leenhardt R, Vasseur P, Li C, Saurin JC, Rahmi G, Cholet F, Becq A, Marteau P, Histace A, Dray X, et al. A neural network algorithm for detection of GI angiectasia during small-bowel capsule endoscopy. Gastrointest Endosc 2019;89(1):189–94.

[16] Aoki T, Yamada A, Aoyama K, Saito H, Tsuboi A, Nakada A, Niikura R, Fujishiro M, Oka S, Ishihara S, et al. Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network. Gastrointest Endosc 2019;89(2):357–63.

[17] Klang E, Barash Y, Margalit RY, Soffer S, Shimon O, Albshesh A, Ben-Horin S, Amitai MM, Eliakim R, Kopylov U. Deep learning algorithms for automated detection of crohn's disease ulcers by video capsule endoscopy. Gastrointest Endosc 2020;91(3):606–13.

[18] Ribeiro T, Mascarenhas M, Afonso J, Cardoso H, Andrade P, Lopes S, Ferreira J, Mascarenhas Saraiva M, Macedo G. Artificial intelligence and colon capsule endoscopy: Automatic detection of ulcers and erosions using a convolutional neural network. J Gastroenterol Hepatol 2022.

[19] Malagelada C, De Iorio F, Azpiroz F, Accarino A, Segui S, Radeva P, Malagelada J-R. New insight into intestinal motor function via noninvasive endoluminal image analysis. Gastroenterology 2008;135(4):1155–62.

[20] Seguí S, Drozdzal M, Pascual G, Radeva P, Malagelada C, Azpiroz F, Vitrià J. Generic feature learning for wireless capsule endoscopy analysis. Comput Biol Med 2016;79:163–72.

[21] Ding Z, Shi H, Zhang H, Meng L, Fan M, Han C, Zhang K, Ming F, Xie X, Liu H, et al. Gastroenterologist-level identification of small-bowel diseases and normal variants by capsule endoscopy using a deep-learning model. Gastroenterology 2019;157(4):1044–54.

[22] Guo X, Yuan Y. Semi-supervised WCE image classification with adaptive aggregated attention. Med Image Anal 2020;64:101733.

[23] Vieira PM, Freitas NR, Lima VB, Costa D, Rolanda C, Lima CS. Multi-pathology detection and lesion localization in WCE videos by using the instance segmentation approach. Artif Intell Med 2021;119:102141.

[24] Adewole S, Fernandez P, Yeghyayan M, Jablonski J, Copland A, Porter MD, Syed S, Brown DE. Lesion2Vec: Deep metric learning for few-shot multiple lesions recognition in wireless capsule endoscopy video. 2021, CoRR, ArXiv:2101.04240.

[25] Pascual G, Laiz P, García A, Wenzek H, Vitrià J, Seguí S. Time-based self-supervised learning for wireless capsule endoscopy. Comput Biol Med 2022;146:105631.

[26] Jain S, Seal A, Ojha A, Yazidi A, Bures J, Tacheci I, Krejcar O. A deep CNN model for anomaly detection and localization in wireless capsule endoscopy images. Comput Biol Med 2021;137:104789.

[27] Smedsrud PH, Thambawita V, Hicks SA, Gjestang H, Nedrejord OO, Næss E, Borgli H, Jha D, Berstad TJD, Eskeland SL, et al. Kvasir-capsule, a video capsule endoscopy dataset. Sci Data 2021;8(1):1–10.

[28] Liang S, Li Y, Srikant R. Enhancing the reliability of out-of-distribution image detection in neural networks. In: International conference on learning representations. 2018.

[29] Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: Proceedings of international conference on learning representations. 2017.

[30] Hendrycks D, Mazeika M, Dietterich T. Deep anomaly detection with outlier exposure. In: International conference on learning representations. 2019, URL https://openreview.net/forum?id=HyxCxhRcY7.

[31] Thulasidasan S, Bhattacharya T, Bilmes J, Chennupati G, Mohd-Yusof J. Combating label noise in deep learning using abstention. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th international conference on machine learning. Proceedings of machine learning research, Vol. 97, PMLR; 2019, p. 6234–43.

[32] Roy AG, Ren J, Azizi S, Loh A, Natarajan V, Mustafa B, Pawlowski N, Freyberg J, Liu Y, Beaver Z, et al. Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. Med Image Anal 2022;75:102274.

[33] Hsu YC, Shen Y, Jin H, Kira Z. Generalized ODIN: Detecting out-of-distribution image without learning from out-of-distribution data. In: 2020 IEEE/CVF conference on computer vision and pattern recognition. CVPR, 2020, p. 10948–57.

[34] Lee K, Lee K, Lee H, Shin J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Proceedings of the 32nd international conference on neural information processing systems. NIPS '18, Red Hook, NY, USA: Curran Associates Inc.; 2018, p. 7167–77.

[35] Li X, Lu Y, Desrosiers C, Liu X. Out-of-distribution detection for skin lesion images with deep isolation forest. In: Machine learning in medical imaging: 11th international workshop, MLMI 2020, held in conjunction with MICCAI 2020, Lima, Peru, october 4, 2020, proceedings 11. Springer; 2020, p. 91–100.

[36] Abati D, Porrello A, Calderara S, Cucchiara R. Latent Space Autoregression for Novelty Detection. In: Proceedings of the IEEE/CVF international conference on computer vision and pattern recognition. 2019.

[37] Lu Y, Xu P. Anomaly detection for skin disease images using variational autoencoder. 2018, ArXiv:1807.01349.

[38] Vyas A, Jammalamadaka N, Zhu X, Das D, Kaul B, Willke TL. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In: Proceedings of the european conference on computer vision. ECCV, 2018, p. 550–64.

[39] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. PMLR; 2019, p. 6105–14.

[40] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 770–8.

[41] Geifman Y, El-Yaniv R. Selectivenet: A deep neural network with an integrated reject option. In: International conference on machine learning. PMLR; 2019, p. 2151–9.