IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 July 2025, accepted 24 July 2025, date of publication 28 July 2025, date of current version 1 August 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3593420

RESEARCH ARTICLE

Using Variational Autoencoders for Out of Distribution Detection in Histological Multiple Instance Learning

FRANCISCO JAVIER SÁEZ-MALDONADO[1], LUZ GARCÍA[2], LEE A. D. COOPER[3,4,5], JEFFERY A. GOLDSTEIN[5], RAFAEL MOLINA[1], (Life Senior Member, IEEE), AND AGGELOS K. KATSAGGELOS[3,6], (Life Fellow, IEEE)

[1]Department of Computer Science and Artificial Intelligence, Universidad de Granada, 18071 Granada, Spain
[2]Department of Signal Theory, Telematics, and Communications, 18071 Granada, Spain
[3]Center for Computational Imaging and Signal Analytics, Northwestern University, Chicago, IL 60611, USA
[4]Chan Zuckerberg Biohub Chicago, Chicago, IL 60642, USA
[5]Department of Pathology, Northwestern University, Chicago, IL 60611, USA
[6]Department of Electrical and Computer Engineering, Northwestern University, Chicago, IL 60611, USA

Corresponding author: Francisco Javier Sáez-Maldonado (fjaviersaezm@ugr.es)

This work was supported by MICIU/AEI/10.13039/501100011033 and by NextGenerationEU/PRTR under Grant PID2022-140189OB-C22 and Grant TED2021-132178B-I00. The work of Lee A. D. Cooper was supported by U.S. National Institutes of Health National Library of Medicine Award under Grant R01LM013523. The work of Jeffery A. Goldstein was supported by U.S. National Institutes of Health under Grant K08EB030120.

ABSTRACT In the context of histological image classification, Multiple Instance Learning (MIL) methods only require labels at Whole Slide Image (WSI) level, effectively reducing the annotation bottleneck. However, for their deployment in real scenarios, they must be able to detect the presence of previously unseen tissues or artifacts, the so-called Out-of-Distribution (OOD) samples. This would allow Computer Assisted Diagnosis systems to flag samples for additional quality or content control. In this work, we propose an OOD-aware probabilistic deep MIL model that combines the latent representation from a variational autoencoder and an attention mechanism. At test time, the latent representations of the instances are used in the classification and OOD detection tasks. We also propose a deterministic version of the model that uses the reconstruction error as OOD score. Panda (prostate tissue) and Camelyon16 (lymph node tissue) are used as train/test in-distribution datasets, obtaining bag classification results competitive with current state-of-the-art models. OOD detection is evaluated performing two experiments for each in-distribution dataset. For Panda, Camelyon16 and ARTIF (prostate tissue contaminated with artifacts) are used as OOD datasets, obtaining 100% AUC in both cases. For Camelyon16, Panda and BCELL (lymph node tissue diagnosed with diffuse large B-cell lymphoma) are used as OOD datasets, obtaining AUCs of 100% and 97%, respectively. Experimental validation demonstrates the models' strong classification performance and effective OOD slide detection, highlighting their clinical potential.

INDEX TERMS Out-of-distribution detection, multiple instance learning, variational autoencoder.

I. INTRODUCTION

Multiple Instance Learning (MIL) is a weakly-supervised learning approach that has recently gained enormous popularity [1], [2]. MIL drastically reduces the annotation effort [3], which is the main bottleneck in many medical Computer Aided Diagnosis (CAD) systems. In MIL, each element in the training set is called a bag, and it is composed of multiple instances. Under the standard MIL assumption [4], each instance has a hidden binary class label, and a bag is positive if, and only if, one or more of its instances are positive. Although different MIL assumptions have been proposed

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Tong.

in the literatures [5] and [6] , the standard assumption of a hidden binary label per instance is the most frequently used [4].

MIL methods are faced with the task of correctly classifying the bag and possibly the instances within the bag while only using bag labels. This is the case of histological image classification, where a frequently sought goal is to determine whether a Whole Slide Image (WSI) contains tumorous tissue [7]. In this case, the WSI is considered the bag, and the instances are small patches from the slide.

There exist two main approaches to designing a MIL classifier: instance-based MIL, where the individual instances are considered to contain the discriminative information for the classification [8]; and embedding-based MIL, where the information extracted from the instances is combined to create a richer representation of the whole bag to be classified. See [9] for a recent and clear presentation of MIL approaches. In practice, embedding-based models have shown superior performance in the classification task. The main reason for this is that aggregating the information from all the instances produces a regularized bag representation which facilitates the classification task [9], [10]. Therefore this approach is the most frequently used in the recent literature.

Most of the state-of-the-art (SOTA) embedding-based deep MIL models utilize an attention mechanism. The first was proposed in [11] and is known as Attention-Based MIL (ABMIL). This model creates permutation-invariant bag representations using the importance of each instance for the classification task. Usually, this results in positive instances in the bag having higher attention values than negative ones, providing an interpretable output of the model. MIL models based on an attention mechanism have evolved a lot since ABMIL was presented, refining their predictive metrics in a variety of ways, such as introducing instance correlations [9], [12], [13], using two branches to further detect key instances [14], [15], or introducing mathematical operators that smooth the attention values along neighbour instances [16], [17].

Although the accuracy of current SOTA deep MIL methods in the classification task is very high, they fail at test time when the input to the model does not have the same structural or morphological features as the training data [18], [19]. In this work, we follow [20], which describes anomaly or Out-of-Distribution (OOD) detection as the process of identifying the samples that do not belong to the training distribution (IN-Distribution, IND).

In digital pathology, detecting OOD samples, either at bag or instance level, is of crucial importance, since flagging a sample as OOD alerts pathologists about the ignorance of the model on the input data. In a real-world scenario, it is common to find slides that contain secondary tumours unseen during training. Tissue cross-contamination also occurs, some instances in the bag come from a different tissue. Furthermore, other artifacts such as blood, folds, or blur can appear [21], [22]. The OOD literature distinguishes between Near and Far OOD problems, which are characterized by their difficulty. Following [23], in Near-OOD, the outlier and inlier classes are highly similar, while in Far-OOD, the outlier is more distinct from the training distribution [24], [25].

The frequent appearance of OOD samples at test time poses an important challenge to MIL models since, to the best of our knowledge, they *know what they know* but, unfortunately, *they don't know what they don't know*. Since they are trained under the closed-world assumption with IND samples, they expect test data to be drawn independently from the same distribution. The main reason for the lack of OOD awareness of current deep MIL methods is that they do not model the underlying data distribution in the training set. Because of this, current MIL models can only use model-agnostic OOD scores like entropy [23] or max-logit [26], which are not trained in the specific data distribution. While the existing literature on the use of MIL in histological image classification continues to grow [1], surprisingly, little attention has been paid to the use of techniques that provide current MIL methods with the ability to model the data distribution.

In this work, we tackle the MIL classification and OOD detection problems by using a deep generative model coupled with a MIL method. To be precise, we use a Variational Autoencoder (VAE) which explicitly models the data distribution and calculates the likelihood of any given instance. The probabilistic latent representations of the instances obtained from the VAE are used in an Attention-Based MIL (ABMIL, [11]) to classify IND bags. Furthermore, those representations are used to compute the marginal likelihood of the instances, which provides the basis for the calculation of a probabilistic OOD score. We name our method VAEABMIL. We also present a deterministic version of VAEABMIL, named DAEABMIL, in which the probabilistic representation is replaced by its deterministic version that is simpler to optimize.

We apply the proposed models to two classification tasks using two well-known datasets: Camelyon16 and Panda. We then present two Far-OOD detection setups, in which the IND and OOD slides do not share the main organ type. Finally, we present two Near-OOD detection experiments using the BCELL and ARTIF datasets (only used as OOD data), in which the IND and OOD slides share the main type of tissue (breast and prostate tissue, respectively). We show that the classification performance of VAEABMIL and DAEABMIL is similar to that of the SOTA deep MIL models for IND data. Furthermore, the OOD detection experiments show that our models excel at detecting OOD samples. This constitutes the main benefit of using VAEABMIL and DAEABMIL: while achieving competitive results in bag-level classification, they are in addition able to determine which bags do not belong to the original IND dataset, which is a task that the rest of the SOTA models are not designed to perform. Our proposals are in fully agreement with [23]: OOD detection is a capability Computer Assisted Diagnosis (CAD) systems should be provided with. We achieve it by making use of the latent representation produced by our models.

In summary, our contributions are the following:

- We introduce VAEABMIL, a novel probabilistic deep MIL method that combines a VAE with ABMIL to

perform IND classification and bag-level OOD detection. We also propose a deterministic version of the model, named DAEABMIL, which shows optimization benefits. VAEABMIL and DAEABMIL constitute the first MIL models with trainable OOD scores.
- We perform an extensive experimentation to validate and show the benefits of our proposal. We use Panda and Camelyon16 as train-test in-distribution datasets. VAEABMIL and DAEABMIL obtain competitive bag classification results with current SOTA MIL models.
- In the OOD detection task, VAEABMIL and DAEABMIL and their respectively tailored OOD scores LOGPX and RECERR, are exhaustively compared against SOTA MIL models using model-agnostic OOD scores. Notice that, so far, no tailored OOD scores have been defined for them. A statistical significance analysis on OOD performance is also included.
- Additionally, we provide a deep analysis of the impact of two different feature extractors in the classification and OOD detection metrics. We experimentally show, for the first time in the OOD-MIL literature, the benefits of using a foundation model for detecting OOD bags in MIL problems.

The rest of the paper is organized as follows. In Section II we first describe the related OOD detection work in digital pathology and then we provide an overview of the ABMIL method and variational autoencoders. In Section III we present VAEABMIL first, then DAEABMIL (Section III-A), and lastly we introduce the proposed OOD scores for both methods (Section III-C). The experiments are presented in Section IV, followed by the conclusions drawn from this work which are explained in Section V. Lastly, further experimental analysis is provided in Appendices A and B.

II. BACKGROUND

In this section, we present the related work (Section II-A). We then mathematically formulate the MIL problem and describe the tools that provide the basis for constructing our MIL method with OOD capabilities (Section II-B).

A. RELATED WORK

The popularity of MIL in digital pathology has grown exponentially due its benefits in WSI classification. See [1], [2], [27] for recent reviews of the SOTA methods.

Out-of-distribution detection undoubtedly plays a very important role in computational pathology [18], [28], [29], reflected by an increasing number of contributions. For instance, [30] provides a comparative analysis of few-shots-exposure and unsupervised uncertainty estimation techniques, proposing a cosine distance-based OOD detection approach for retinal OCT images. Notice that other reconstruction errors can also be used [31]. Deterministic uncertainty estimations of classifiers and ensembles for OOD threshold-based detection are presented in [32] in the framework of breast and prostate cancer detection in histopathological images. Furthermore, a probabilistic uncertainty estimation is proposed in [33] using a Bayesian U-net to detect anomalies in OCT images. In [21], a deep kernel model is used to detect histological artifacts, blur, and folds in glass slides of bladder tumour resections. Lastly, the most recent works review the use of AnoDDPM [34] and AnoLDM [35] for OOD detection in digital pathology. However, none of the previously mentioned works are developed under the MIL framework.

The importance of using a good latent representation of the data has been widely acknowledged in different research areas, see, for instance [36] and [37]. VAEs provide a good example of it and they have been frequently used for standalone OOD detection [38], [39], [40]. In the medical domain, they have been used for unsupervised anomaly localization in CT scans [41] or anomaly detection in electrocardiogram records [42], always outside the MIL paradigm. In [43], a VAE is used to define a MIL model that creates a disentangled representation of the instance features, later used for OOD generalization: the task in which the model is used on samples from another dataset and is expected to maintain its classification performance. Note that OOD generalization is not the same task as OOD detection, so [43] does not propose a MIL based OOD model. Thus, the use of VAEs for OOD detection in the MIL framework remains, so far, unexplored.

The aggregation of instance-level OOD scores to perform OOD detection is explored in [23], where multiple patch-level CNNs are trained and the patch-level entropy is aggregated to obtain a bag-level OOD score. Notice that this is not a MIL classification model but the use of the estimated patch-level classification probabilities to define a bag-level OOD score.

To conclude this section, we remark that although recent references on the use of OOD detection methods in WSI classification exist, none of them has been formulated using the MIL paradigm. Our VAEABMIL and DAEABMIL constitute pioneering approaches on providing MIL methods with OOD capabilities.

B. DEEP MULTIPLE INSTANCE LEARNING

Our work focuses on embedding OOD detection capabilities in deep MIL classification models. For this reason, we start by presenting the elements of the MIL setup for the classification task. In MIL, each element of the dataset is a pair $(\mathbf{X}, y)$, where $\mathbf{X} \in \mathbb{R}^{N_b \times P}$ is a bag with $P \in \mathbb{N}$ the dimension of the feature space provided by a pretrained encoder, and y is the bag label. Each bag is composed of N_b instances, $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_{N_b}]^T$. In this work, we consider a binary classification problem. Following the so-called standard MIL assumption [4], a bag is positive if, and only if, at least one of its instances is positive. That is, $y = \max\{y_i\}_{i=1}^{N_b} \in \{0, 1\}$, where y_i is the label of the instance $\mathbf{x}_i$. We consider our dataset to have B pairs, and we will use the notation $\mathbf{X}^b$ to denote the $b-$th bag of the dataset, with instances $\mathbf{x}_1^b, \cdots, \mathbf{x}_{N_b}^b$. Unless necessary, we will omit the bag reference (b) for simplicity.

The goal in the classification task is to learn a function that maps each input bag to a label. At test time, a previously unseen bag $\mathbf{X}^\star$ is received by the model which outputs a class for it. As indicated in the introduction, we will follow the *embedding-based* approach to designing a MIL classifier that solves the standard MIL problem. The model creates a representation $\mathbf{B}$ of the bag by aggregating the information of its instances and uses it to assign a label to each bag. To create the aforementioned representation, the current most relevant models are *deep attention* MIL models. These methods are composed of three main blocks: a feature refiner, an attention mechanism and a classifier. We now describe each of the blocks individually.

First, in deep attention MIL models, each instance $\mathbf{x}_i$ of the bag is processed using a neural network g_η, the feature refiner, with parameters η. This creates a latent representation of that instance $\mathbf{z}_i = g_\eta(\mathbf{x}_i) \in \mathbb{R}^D$, with $D \in \mathbb{N}$ the latent space dimension, which contains its most relevant information. We denote by $\mathbf{Z}^T = [\mathbf{z}_1, \cdots, \mathbf{z}_{N_b}]$ the matrix containing the latent representations of the instances in a bag.

In embedding-based MIL, the information of the instances is aggregated to create a richer representation of the whole bag that takes into account *how important each instance is in the bag representation*. This importance value is often called *attention value*, and it is widely used in many current deep MIL models such as ABMIL [11], TRANSMIL [12] or DTFDMIL [9]. In this work, we build upon the well known ABMIL model, in which the attention module computes the vector of attention values $\mathbf{f}$ as follows: considering $\mathbf{W} \in \mathbb{R}^{L\times D}$ and $\mathbf{w} \in \mathbb{R}^L$ to be learnable weights and $L \in \mathbb{N}$,

$$\mathbf{F}_{\text{mid}} = \tanh(\mathbf{Z}\mathbf{W}^T) \in \mathbb{R}^{N_b \times L} \tag{1}$$

$$\mathbf{f} = \mathbf{F}_{\text{mid}}\mathbf{w} \in \mathbb{R}^{N_b}. \tag{2}$$

The softmax is applied to $\mathbf{f}$ to obtain the attention values that are all positive and add up to one. Then, each obtained value is multiplied by its corresponding embedding and aggregated to obtain the final bag representation $\mathbf{B}$ as:

$$\mathbf{B} := \mathbf{Z}^T \text{Softmax}(\mathbf{f}) \in \mathbb{R}^D \tag{3}$$

This bag representation aggregates the information of the instances of the bag according to their importance in the classification task. Finally, we pass it through a simple linear classifier $c_\gamma : \mathbb{R}^D \to [0, 1]$ with parameters γ, which assigns to each bag its probability of being of the positive class.

C. VARIATIONAL AUTOENCODERS

The usage of VAEs in MIL is the key proposal of our work. In VAEs, instead of considering a single, deterministic latent representation $\mathbf{z}$ for each input, they place a prior distribution $p(\mathbf{z})$ over that latent encoded representation. Given $\mathbf{z}$, a probabilistic reconstruction is obtained using an observation model $p(\mathbf{x} \mid \mathbf{z})$. Typically, the prior is chosen to be a standard Gaussian distribution $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ since it enforces smoothness, as well as beneficial structural and continuity properties in the latent space. The observation model is also chosen Gaussian $p_\theta(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(\mathbf{x} \mid m_\theta(\mathbf{z}), \sigma^2_\theta(\mathbf{z})\mathbf{I})$, where the mean function $m_\theta(\mathbf{z})$ and covariance $\sigma^2_\theta(\mathbf{z})\mathbf{I}$ are parameterized by neural networks with parameters θ.

With this selection of the prior and likelihood distributions, predictions in VAEs are made by integrating over the posterior distribution $p(\mathbf{z} \mid \mathbf{x})$ which, unfortunately, can not be computed in closed form. For this reason, Variational Inference (VI) [44] is often used as a form of approximating the exact posterior using a Gaussian variational distribution $q_\phi(\mathbf{z} \mid \mathbf{x})$

$$p(\mathbf{z} \mid \mathbf{x}) \approx q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\mathbf{z} \mid m_\phi(\mathbf{x}), \sigma^2_\phi(\mathbf{x})\mathbf{I}), \tag{4}$$

where the mean and the covariance are parameterized by neural networks ($m_\phi(\mathbf{x})$ and $\sigma^2_\phi(\mathbf{x})$, respectively) with parameters ϕ. To optimize the parameters of the likelihood and posterior distributions we maximize the Evidence Lower Bound (ELBO) [45], which lower bounds the marginal likelihood of the data $p(\mathbf{x})$. The ELBO in VAEs for a sample $\mathbf{x}$ takes the form:

$$\mathcal{L}^{\text{VAE}}_{\phi,\theta}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] - KL(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p(\mathbf{z})), \tag{5}$$

which can be optimized via Monte-Carlo Sampling [44].

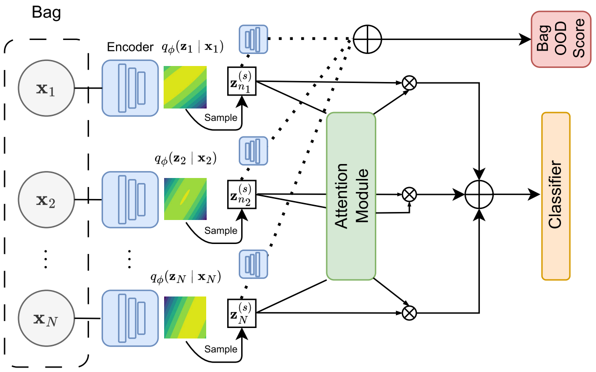


FIGURE 1. Graphical overview of the structure of VAEABMIL. Each instance $\mathbf{x}_i$ is encoded to obtain its approximated posterior distribution $q(\mathbf{z}_i \mid \mathbf{x}_i)$ using the encoder of the VAE. Then, a sample $\mathbf{z}_i(s) \sim q(\mathbf{z}_i \mid \mathbf{x}_i)$ is obtained, which is used both in the classification and OOD detection tasks. The classification is done using the Attention MIL paradigm on the samples from the approximated posterior. The OOD detection is performed using the decoder of the VAE.

III. PROPOSED METHODS

In this section, we propose a novel deep MIL model with OOD capabilities named VAEABMIL, built upon a VAE, described in Section II-C and the attention mechanism described in Section II-B. The use of a VAE is motivated by the need to model the data distribution in order to detect possible OOD bags that may appear in the test set. The attention mechanism in ABMIL is used since it is the base of current SOTA MIL models. In VAEABMIL, instead of using the deterministic latent embedding $\mathbf{z}$ (with no OOD capabilities) used in ABMIL, we make use of a VAE which will replace the MIL feature refiner g_η and will be equipped with OOD capabilities. Notice that this is a main novelty and an important benefit of VAEABMIL: it is a deep MIL model capable of both classifying bags and also detecting OOD samples. For the embedded VAE, we use the typical Gaussian observation and prior models presented in Section II-C, which will allow us to define a probabilistic OOD score (see Section III-C). Let us now provide the mathematical formulation.

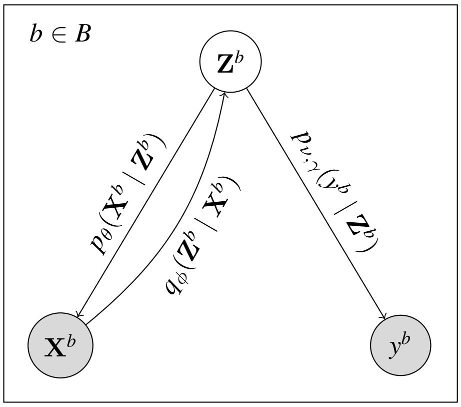

$b \in B$
$\mathbf{Z}^b$
$p_\theta(\mathbf{X}^b \mid \mathbf{Z}^b)$
$q_\phi(\mathbf{Z}^b \mid \mathbf{X}^b)$
$p_{\nu,\gamma}(y^b \mid \mathbf{Z}^b)$
$\mathbf{X}^b$
y^b

FIGURE 2. **Probabilistic graphical depiction of VAEABMIL. Given the latent variables $\mathbf{Z}^b$, the bag label y^b is independent of the observed bag $\mathbf{X}^b$. We use $q_\phi(\mathbf{Z}^b \mid \mathbf{X}^b)$ for both the classification and OOD detection tasks.**

Given an observed bag $\mathbf{X}^b$ and denoting by $\mathbf{Z}^b$ the associated bag of *random* latent representations of its instances, each $\mathbf{z}_i^b$ is responsible for the probabilistic generation of $\mathbf{x}_i^b$, $i = 1, \ldots, N_b$ using the VAE formulation described in Section II-C. We then make $\mathbf{Z}^b$ solely responsible for the MIL classification of the bag, that is, $\mathbf{X}^b$ and y^b are conditionally independent given $\mathbf{Z}^b$. We further use the attention mechanism in Equation (2) and the weighted-by-attention average of the instances in Equation (3) to obtain a bag representation $\mathbf{B}^b$ that summarizes the information of the instances. Using the bag representation, we can compute the probability of the bag label $p_{\nu,\gamma}(y^b \mid \mathbf{Z}^b) = \text{Bern}(c_\gamma(\mathbf{B}^b))$, with $\nu = \{\mathbf{W}, \mathbf{w}\}$. A complete overview of the model can be observed in Figure 1. Also, the corresponding probabilistic graphical model is displayed in Figure 2. Letting $\mathbb{X} = \{\mathbf{X}^1, \cdots, \mathbf{X}^B\}$, $\mathbb{Y} = \{y^1, \cdots, y^B\}$ and $\mathbb{Z} = \{\mathbf{Z}^1, \cdots, \mathbf{Z}^B\}$, the joint distribution takes the form:

$$p_{\theta,\nu,\gamma}(\mathbb{Y}, \mathbb{X}, \mathbb{Z}) = p_{\nu,\gamma}(\mathbb{Y} \mid \mathbb{Z}) p_\theta(\mathbb{X} \mid \mathbb{Z}) p(\mathbb{Z})$$
$$= \prod_{b=1}^{B} \left(\underbrace{p_{\nu,\gamma}(y^b \mid \mathbf{Z}^b)}_{\text{Classification likelihood}} \prod_{i=1}^{N_b} \left(\underbrace{p_\theta(\mathbf{x}_i^b \mid \mathbf{z}_i^b)}_{\text{VAE likelihood}} p(\mathbf{z}_i^b) \right) \right), \tag{6}$$

where we have used the assumption of bag-level factorization in the classification-likelihood term and the instance-level factorization in the VAE-likelihood term. Notice that, for each b, $\mathbf{Y}^b$ and $\mathbf{X}^b$ are independent given $\mathbf{Z}^b$ but they become dependent when $\mathbf{Z}^b$ is integrated on. This makes the unsupervised representation of the patches and the MIL classification dependent tasks. Note that, by removing the randomness on $\mathbf{Z}^b$ and ignoring the decoder, we obtain the standard ABMIL. To make predictions, the latent variables $\mathbf{Z}^b$ are marginalized using the posterior distribution $p(\mathbf{Z}^b|\mathbf{X}^b)$. Unfortunately, this distribution cannot be calculated in closed form and so we follow the procedure in VAEs, resorting to a variational approximation that factorizes across bags and instances. This variational posterior distribution takes the form:

$$q_\phi(\mathbb{Z}|\mathbb{X}) = \prod_{b=1}^{B} q_\phi(\mathbf{Z}^b|\mathbf{X}^b) = \prod_{b=1}^{B}\prod_{i=1}^{N_b} q_\phi(\mathbf{z}_i^b \mid \mathbf{x}_i^b)$$
$$= \prod_{b=1}^{B}\prod_{i=1}^{N_b} \mathcal{N}(\mathbf{z}_i^b \mid m_\phi(\mathbf{x}_i^b), \sigma_\phi^2(\mathbf{x}_i^b)\mathbf{I}), \tag{7}$$

where each m_ϕ and σ_ϕ^2 are the ones defined for the VAE (see Section II-C). Notice the simplification in the isotropic structure of the posterior covariance approximation for computational reasons, since the covariance matrix size scales quadratically with the number of instances in a bag, which can be very large depending on the patch and WSI sizes. Using more complex posteriors would drastically increase the optimization complexity of the model. We optimize the parameters of our model, $\phi, \theta, \nu, \gamma$, by maximizing the ELBO (or, equivalently, minimizing the minus ELBO), which in the proposed model takes the form:

$$\mathcal{L}_{\phi,\theta,\nu,\gamma}^{\text{VAEABMIL}}(\mathbb{X}, \mathbb{Y}) = \mathbb{E}_{q_\phi(\mathbb{Z}|\mathbb{X})}\left[\log \frac{p_{\theta,\nu,\gamma}(\mathbb{Y}, \mathbb{X}, \mathbb{Z})}{q_\phi(\mathbb{Z}|\mathbb{X})}\right] \tag{8}$$
$$= \mathbb{E}_{q_\phi(\mathbb{Z}|\mathbb{X})}$$
$$\left[\log \frac{\prod_{b=1}^{B}\left(p_{\nu,\gamma}(y^b \mid \mathbf{Z}^b)\prod_{i=1}^{N_b}\left(p_\theta(\mathbf{x}_i^b \mid \mathbf{z}_i^b)p(\mathbf{z}_i^b)\right)\right)}{\prod_{b=1}^{B}\prod_{i=1}^{N_b} q_\phi(\mathbf{z}_i^b \mid \mathbf{x}_i^b)}\right]$$
$$= \sum_{b=1}^{B}\Big[\mathbb{E}_{q_\phi(\mathbf{Z}^b|\mathbf{X}^b)}[\log p_{\nu,\gamma}(y^b \mid \mathbf{Z}^b)] \tag{9}$$
$$+ \sum_{i=1}^{N_b} \mathbb{E}_{q_\phi(\mathbf{z}_i^b|\mathbf{x}_i^b)}\left[\log p_\theta(\mathbf{x}_i^b \mid \mathbf{z}_i^b)\right] \tag{10}$$
$$- \sum_{i=1}^{N_b} KL\big(q_\phi(\mathbf{z}_i^b \mid \mathbf{x}_i^b) \parallel p(\mathbf{z}_i^b)\big)\Big]. \tag{11}$$

The term in (9) is the classification log likelihood, which explains how well the model classifies the bags. The VAE log likelihood, Equation (10), measures the quality of the instance reconstruction of the VAE. The last term, (11) is the Kullback-Leibler (KL) divergence between the variational posterior and the Gaussian prior, which aims to regularize the variational posterior. The last two terms together are responsible for the OOD detection and the learning of the manifold of the IND data. Notice that the KL divergence is crucial to maintain the properties of the latent space [46], therefore no term can be suppressed from this loss in order to maintain the performance of the model in both classification and OOD detection tasks.

A. A DETERMINISTIC VERSION OF VAEABMIL

Although the presented probabilistic model VAEABMIL is theoretically sound, it is known that probabilistic models are harder to optimize than deterministic ones. This provides the motivation to derive DAEABMIL, a deterministic version of VAEABMIL. To achieve this, we restrict the posterior distribution of VAEABMIL in Equation (4) to be a Dirac's delta $\delta(\mathbf{z} - m_\phi(\mathbf{x}))$. Then, the instance latent representations $\mathbf{Z}$ become unique, rather than random variables. The loss function for DAEABMIL then becomes:

$$\mathcal{L}^{\text{DAEABMIL}}_{\phi,\theta,\nu,\gamma}(\mathbb{X}, \mathbb{Y}) = \sum_{b=1}^{B} \Big(\mu \log p_{\nu,\gamma}(y^b \mid \mathbf{Z}^b) + \alpha \sum_{i}^{N_b} \left\| \mathbf{x}_i^b - m_\theta(\mathbf{z}_i^b) \right\|^2 + \beta \sum_{i}^{N_b} \left\| \mathbf{z}_i^b \right\|^2 \Big), \tag{12}$$

where $m_\theta(\mathbf{z}_i^b)$ is the decoding of $\mathbf{z}_i^b$ and μ, α, β are positive and add up to one. What is more, this model generalizes ABMIL, since taking $\alpha = \beta = 0$ ABMIL is recovered. Notice here, as we did with VAEABMIL, that the last two terms together are responsible for the OOD detection and the learning of the manifold of the training IND data. This manifold is now deterministic. With DAEABMIL we obtain faster inference, but it loses the probabilistic prediction.

B. IND CLASSIFICATION PREDICTIONS

In VAEABMIL, to make classification predictions on new test bags, we use the latent variables generated by the VAE. Given a test bag $\mathbf{X}^\star = [\mathbf{x}_1, \cdots, \mathbf{x}_{N_\star}]$ with $N_\star$ instances, we define $\mathbf{Z}^\star_{(s)} = [\mathbf{z}_1^{\star,(s)}, \cdots, \mathbf{z}_{N_\star}^{\star,(s)}]$, where $\mathbf{z}_i^{\star,(s)} \sim q_\phi(\mathbf{z}_i^\star \mid \mathbf{x}_i^\star)$ is a sample from the approximated posterior of instance $\mathbf{x}_i^\star$. Then, we approximate the predictive distribution using S Monte Carlo samples as:

$$\begin{aligned} p_{\nu,\gamma}(y^\star \mid \mathbf{X}^\star) &= \int p_{\nu,\gamma}(y^\star \mid \mathbf{Z}^\star) p(\mathbf{Z}^\star | \mathbf{X}^\star) d\mathbf{Z}^\star \\ &\approx \int p_{\nu,\gamma}(y^\star \mid \mathbf{Z}^\star) q_\phi(\mathbf{Z}^\star | \mathbf{X}^\star) d\mathbf{Z}^\star \\ &\approx \frac{1}{S} \sum_{s} p_{\nu,\gamma}(y^\star \mid \mathbf{Z}^\star_{(s)}), \end{aligned} \tag{13}$$

where, in the first equality, we have used the conditional independence of $y^\star$ and $\mathbf{X}^\star$ given $\mathbf{Z}^\star$.

In the case of DAEABMIL, the instance embeddings are deterministic, so classification predictions are obtained as in ABMIL (see Section II-B), with the significative difference that the latent embedding space was also trained to be robust to instance-level reconstruction (and, thus, capable to detect OOD samples) using a deterministic autoencoder.

C. OUT-OF-DISTRIBUTION DETECTION

One of the most important advantages of VAEABMIL and DAEABMIL is their capability to model the instance-level data distribution $p(\mathbf{x})$ and, hence, detect OOD bag samples. In this work, we propose to use an aggregation of the instance-level log marginal likelihood as the bag-level OOD score. This score is motivated by the probabilistic meaning of the marginal likelihood: the lower marginal likelihood of $\mathbf{x}$, the higher probability of $\mathbf{x}$ being OOD.

To calculate this score, for each instance $\mathbf{x}_i^\star$ we first consider $\mathbf{z}_i^\star$, the unsupervised random representation of $\mathbf{x}_i^\star$, to compute the marginal distribution $p(\mathbf{x}_i^\star)$ which can be obtained using importance sampling with S Monte Carlo samples as:

$$\begin{aligned} p(\mathbf{x}_i^\star) &= \int \frac{p_\theta(\mathbf{x}_i^\star \mid \mathbf{z}_i^\star) p(\mathbf{z}_i^\star)}{q_\phi(\mathbf{z}_i^\star \mid \mathbf{x}_i^\star)} q_\phi(\mathbf{z}_i^\star \mid \mathbf{x}_i^\star)\, d\mathbf{z}_i^\star \\ &= \mathbb{E}_{q_\phi(\mathbf{z}_i^\star | \mathbf{x}_i^\star)} \left[\frac{p_\theta(\mathbf{x}_i^\star \mid \mathbf{z}_i^\star) p(\mathbf{z}_i^\star)}{q_\phi(\mathbf{z}_i^\star \mid \mathbf{x}_i^\star)} \right] \\ &\approx \frac{1}{S} \sum_{s=1}^{S} \frac{p_\theta(\mathbf{x}_i^\star \mid \mathbf{z}_i^{\star,(s)}) p(\mathbf{z}_i^{\star,(s)})}{q_\phi(\mathbf{z}_i^{\star,(s)} \mid \mathbf{x}_i^\star)} \end{aligned} \tag{14}$$

This marginal distribution indicates how likely it is that a sample belongs to the training data distribution. We aggregate the instance-level score using the mean to compute VAEABMIL's bag-level OOD score as:

$$\text{LOGPX}(\mathbf{X}^\star) := \frac{1}{N_\star} \sum_{i=1}^{N_\star} \left\{ - \log p\left(\mathbf{x}_i^\star\right) \right\}. \tag{15}$$

The higher the LOGPX score, the more likely the bag is OOD. Algorithmically, given a test bag $\mathbf{X}^\star$, the posterior distribution $q_\phi(\mathbf{z}_i^\star \mid \mathbf{x}_i^\star)$ of each of its instances is computed using Equation (4). Then, we sample S times from the approximated posterior of each instance, obtaining $\mathbf{z}_i^{\star,(s)}$ for $i = 1, \cdots, N_\star$ and $(s) = 1, \cdots, S$. We then use Equation (14) to obtain an approximation of the marginal likelihood of each instance. Lastly the instance-level scores are aggregated using Equation (15). As a note, other aggregations (such as the maximum of the minus log marginal likelihoods) could be considered, but we have found the mean to be the best in practice (see the results with the MAX aggregation in Appendix B).

In the case of DAEABMIL, we can not compute the log marginal likelihood since the model is no longer probabilistic. However, we can compare the reconstruction with the original sample to see if the deterministic autoencoder can accurately reconstruct it. Since we expect OOD samples to have higher reconstruction errors, we propose to use the mean of the reconstruction errors as DAEABMIL's bag-level OOD score:

$$\text{RECERR}(\mathbf{X}^\star) := \frac{1}{N_\star} \sum_{i=1}^{N_\star} \left\| \mathbf{x}_i^\star - m_\phi(\mathbf{z}_i^\star) \right\|^2. \tag{16}$$

As in the previous case, higher reconstruction errors indicate a higher chance of a sample being OOD. Algorithmically, given a test bag $\mathbf{X}^\star$ we first compute its deterministic latent representation $\mathbf{Z}^\star$. Then, we reconstruct each instance using the decoder $m_\theta(\mathbf{z})$ and the bag-level OOD score is obtained using Equation (16).

Interestingly, we have provided our model with prediction and OOD detection capabilities. We have constrained the latent representations to be useful for the classification task but also for the OOD detection task. The defined OOD scores reflect the probability that an input belongs to the training distribution. Thus, they provide reliable and interpretable, tailored to the data, OOD scores in MIL settings. Recall that existing MIL models must rely on model-agnostic OOD scores"metrics derived from the model's output rather than the underlying data distribution" to estimate the likelihood of an input being OOD.

IV. EXPERIMENTS

In this section we first describe the datasets, the experimental methodology, and the models used. This description is followed by the results supported by figures and graphical tables. A discussion of the limitations of the proposed approach concludes the section.

A. DATASETS

Four different datasets are used to validate our proposed approach. The Camelyon16 (CAM16) dataset [47] is used to address the task of detecting metastases in hematoxylin and eosin (H&E) stained WSIs of breast cancer metastases. It is composed of 270 training and 130 test images. This dataset is public and it was presented in the Camelyon16 Grand Challenge.[1]

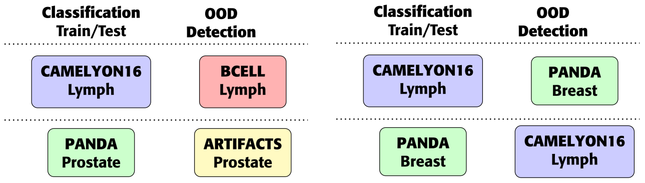
The Panda (PANDA) dataset [48] is a public dataset that was presented in the Panda Grand Challenge[2]. It contains prostate tissue WSIs. Here we use it for a binary cancer-no cancer classification problem. In total, PANDA contains 8822 training slides and 1794 test slides.

Studies of sentinel lymph node biopsies for breast cancer show that 1.6% contain lymphoma. The third dataset (BCELL) contain 26 lymph node tissue WSIs diagnosed with diffuse large B-cell lymphoma. Here it will be used as OOD data.

The fourth dataset (ARTIF) contains 27 prostate tissue slides with different types of artifacts such as blur, foreign tissue or technical artifacts. Here it will be used as OOD data. This is a very interesting challenge since artifacts commonly appear in real-world clinical scenarios.

The datasets have been carefully selected to offer a great variability of scenarios. Two different main tissue types are used for training (breast CAM16 and prostate in PANDA). Those datasets also present differences in the size of their WSIs and, hence, in the average number of extracted patches per slide. In the experiments BCELL will be used as OOD for CAM16 since they both contain lymph node tissue. ARTIF will be used as OOD for PANDA. Both contain prostate WSIs.

The four datasets are processed as follows. For each image, 512 × 512 pixel patches (instance) are extracted with the highest available resolution. The provided masks in CAM16 and PANDA are used to produce bag labels while instances remain unlabelled. Since prostate tissue biopsies (PANDA) are smaller than lymph node sections (CAM16), PANDA bags contain on average a smaller number of instances. Patch features are then extracted utilizing two different pre-trained models: Resnet50 with Barlow Twins (BT) self-supervised learning [49], using the weights provided in [50], and the general-purpose self-supervised foundation model for pathology UNI [51]. UNI was trained using more than 100 million images across 20 major tissue types. The usage of these two feature extractors allows the analysis of their influence in the classification and OOD detection tasks. Patches and feature extraction is performed using the code from CLAM [52].[3]



Near OOD experiments. Far OOD experiments.

FIGURE 3. Graphical description of Near and Far OOD experiments. The main tissue type is indicated under the dataset name. Each experiment is performed using two feature extractors (UNI and BT).

B. EXPERIMENTAL DESIGN

In this work, we assume that detecting OOD samples is a test-time task. That is, our training (IND) datasets will be free from OOD data, and the OOD samples will appear during testing. Thus, to evaluate each model, each experiment consists of two different steps:

1) Classification step, where each model is trained on IND datasets (CAM16 and PANDA, independently), and evaluated in the IND test set.
2) OOD detection step, where we use the already trained models to measure their OOD detection performance using an OOD dataset.

For the classification task, the proposed models are compared with the following five SOTA MIL models: DTFDMIL [15] which uses pseudo bags to create a double-tier MIL with distilled bag features, TRANSMIL [12], which uses a Transformer architecture to create bag representations which take into account instance correlations, DSMIL [14], which uses instance correlations adding a pyramidal fusion of WSI features, and CLAM [52], which uses multiple attention branches for each class. Lastly, we also use the baseline ABMIL [11].

In the OOD detection task, we create four pairs of datasets (IND data, OOD data): (CAM16, BCELL), (PANDA, ARTIF), (CAM16, PANDA), (PANDA, CAM16) . In the first two pairs, (CAM16, BCELL), (PANDA, ARTIF), IND and OOD slides share the main tissue. Therefore these experiments are defined as *Near OOD* detection scenarios, representing a harder OOD task due to the similarity of the tissues present in the slides. The

[1] Link to Camelyon16's challenge.
[2] Link to Panda's challenge.
[3] CLAM's code in GitHub.

other two pairs are considered a *Far* OOD detection problem. A summary of these experiments can be found in Figure 3.

To perform the OOD detection task, bag-level OOD scores are computed. LOGPX (Equation (15)) and RECERR (Equation (16)) are respectively proposed for VAEABMIL and DAEABMIL. For the models we compare against, since they are not designed to handle IND/OOD discrimination, we resort to post-hoc OOD scores. Using the model logits ℓ, we compute the Maximum Logit Score (MLS) [26] and the Entropy of the prediction [23], which (for a two class problem) takes the form

$$\mathrm{H}(p) = -\left(p \log p + (1-p)\log(1-p)\right), \qquad (17)$$

with $p = \mathrm{sigmoid}(\ell)$. Entropy and MLS scores are also computed for VAEABMIL and DAEABMIL. In this section, we report, for each model, the highest metric value obtained across all the OOD scores. In Appendix B, we provide complete results for all models with the different OOD scoring methods. Notice that other model-agnostic OOD scores could be selected, but Entropy and MLS are the most frequently used in the OOD literature.

To compare the results, AUC [53] is used. It quantifies a model's ability to distinguish between positive and negative classes across all possible classification thresholds. In the classification task, model logits are used to compute the AUC. In the OOD detection task, the AUC is computed based on the OOD scores obtained for each model.

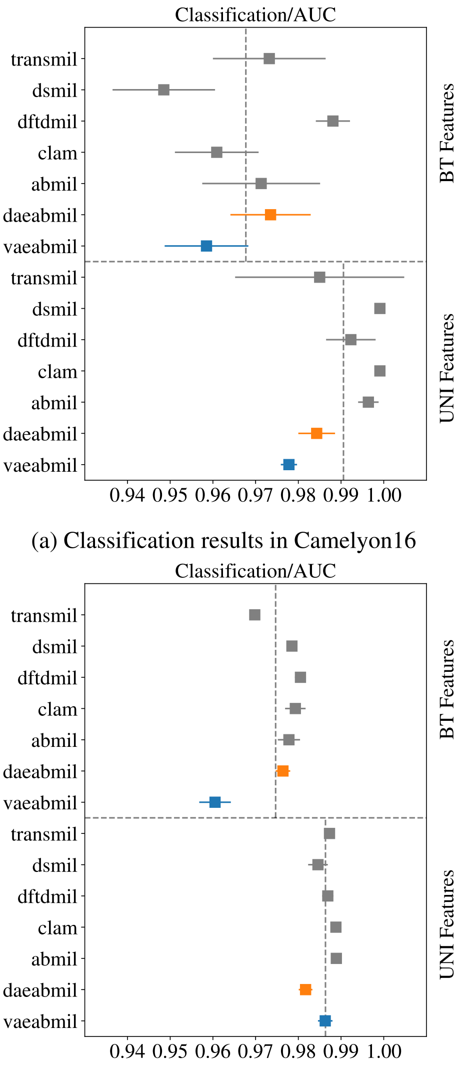
C. IMPLEMENTATION DETAILS

Each model is run three times with different train/validation splits to provide statistically reliable results. We split 20% of the train set and used it as the validation set. We train each model for 100 epochs in CAM16 and 50 in PANDA with no early stopping, using a learning rate of 10^{-4} for all the models but TRANSMIL, for which we use 10^{-5}. For each run, test metrics are computed using the model weights corresponding to the highest validation AUC achieved during training. We code the models using *Pytorch* [54], and we use the Adam optimizer [55].

For the architecture of VAEABMIL and DAEABMIL, in both cases we use simple autoencoders composed of three linear layers with sizes $[512, 256, 128]$ as the encoder and we utilize the same dimensions for the decoder. In DAEABMIL, we use $\mu = 1$, and $\alpha = \beta = 0.3$ in Eq. (12) to train the model. To predict the variances in VAEABMIL, we produce a single value that is used across all the latent dimensions and use $S = 1$ Monte Carlo sample for inference. The models are trained in a single Nvidia 3090 GPU with 24 gigabytes of RAM. The rest of the model follows the implementation of the original ABMIL. The code is available at https://github.com/fjsaezm/VAEABMIL.

D. CLASSIFICATION RESULTS

Figure 4 shows the AUC metric in the bag classification task for CAM16 and PANDA, using both BT and UNI feature extractors.

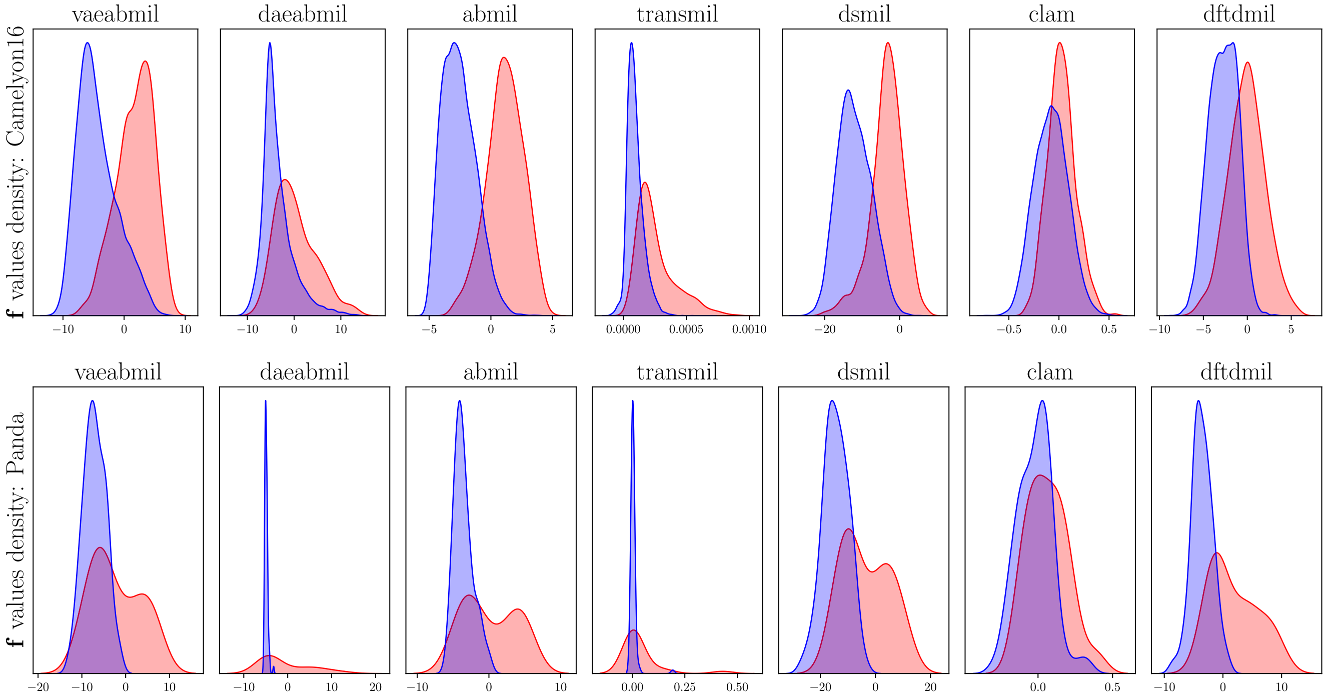


(a) Classification results in Camelyon16

(b) Classification results in Panda

FIGURE 4. Classification results for both CAM16 and PANDA datasets. The presented metric is the test AUC(right is better). Mean and standard deviations are reported for each model. The results with both feature extractors are separated by the horizontal dashed line. The vertical, dashed lines represent the mean performance of the models using each feature extractor.

For the CAM16 dataset, Figure 4a shows that the models are, regardless the feature extractor, very accurate for this benign/malignant classification, with the worst performance being better than 0.95 AUC. With both feature extractors, VAEABMIL and DAEABMIL perform similarly to the rest of the models. In the cases where our models perform worse than the rest, the highest difference in AUC does not exceed 1%. This is compensated by their additional OOD detection capabilities. Comparing the results across the different feature extractors, the models perform clearly better when using UNI. This is observed in the vertical dashed black lines, which represent the average of the means of all the models using the corresponding features. This indicates that UNI produces excellent features of the patches, facilitating the classification task.



(a) Approximated densities of the predicted unnormalized attention values **f** in positive slides from CAM16 and PANDA, using UNI features.

Dataset	VAEABMIL	DAEABMIL	ABMIL	CLAM	DTFDMIL	DSMIL	TRANSMIL
CAM16	$0.929_{0.018}$	$0.846_{0.063}$	$0.976_{0.009}$	$0.980_{0.006}$	$0.980_{0.005}$	$\mathbf{0.987}_{0.001}$	$0.653_{0.069}$
PANDA	$0.767_{0.002}$	$0.774_{0.012}$	$0.805_{0.002}$	$0.798_{0.010}$	$\mathbf{0.822}_{0.006}$	$0.804_{0.002}$	$0.620_{0.060}$

(b) Instance-Level AUC (using the unnormalized attention values) obtained by all the models using UNI features in both CAM16 and PANDA datasets. Mean and standard deviations are reported.

FIGURE 5. Instance-level results, using the unnormalized attention values f and UNI features.

Figure 4b presents the classification results on the PANDA dataset, which show trends similar to those observed in CAM16. When using BT features, VAEABMIL performs approximately 2% worse than the other models, whereas DAEABMIL performs comparably to the SOTA methods. This 2% performance gap between VAEABMIL and DAEABMIL is also observed in CAM16, highlighting the optimization advantages of DAEABMIL over VAEABMIL for classification tasks using BT features. In contrast, when using UNI features, all models achieve near-perfect classification performance, with AUC scores exceeding 0.98.

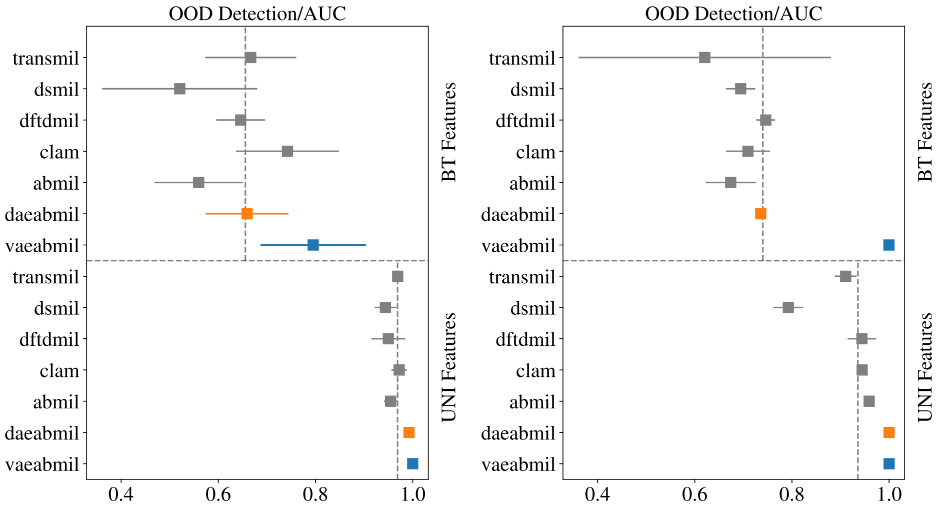
To conclude this section, we compare the attention values (Equation (2)), provided by each classifier. Figure 5a shows the instance-level attention prediction in positive bags in both CAM16 and PANDA, using UNI features. Visually, VAEABMIL performs better than DAEABMIL in CAM16 and equally in PANDA, emphasizing the benefits of obtaining a probabilistic, continuous latent space. This is confirmed by the quantitative results shown in Table 5b. Compared with the rest of the models (except for TRANSMIL), we observe that the proposed models perform slightly worse. However, as it was shown in Figure 4, the bag-level performance of our methods is similar to that of the rest of the models. Notice that, similarly, TRANSMIL obtains poor attention values but high bag-level classification metrics.

E. FAR OOD DETECTION

OOD detection results are now presented, starting with Far OOD experiments, where IND and OOD data do not share the main tissue type and, therefore, we expect an easier task. Results shown in this section are supported by the statistical significance analysis performed in Appendix A.

1) (CAM16, PANDA)

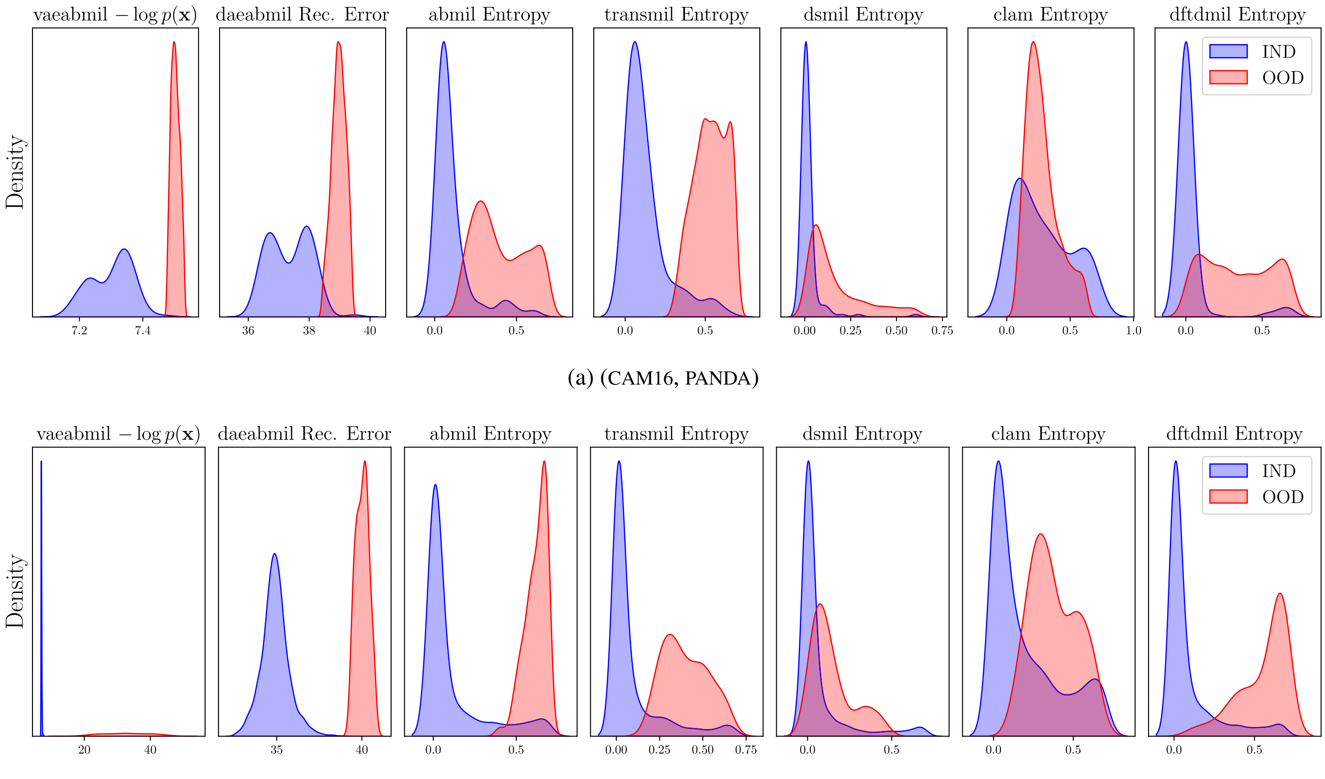
Models trained with CAM16 (see the classification performance in Section IV-D) are now evaluated using PANDA as OOD dataset. Figure 6a shows that, when BT features are used, VAEABMIL obtains the best OOD detection result, and DAEABMIL is on average with the rest of the models. Such behaviour is caused by two main reasons: a) The difficulties in the two-task optimization process which our proposal suffers from (specified in Section IV-G), and b) the deterministic latent space in DAEABMIL might not be flexible enough to produce far-apart representations for the CAM16 and



(a) (CAM16, PANDA)

(b) (PANDA, CAM16)

FIGURE 6. Far OOD detection results. The presented metric is the AUC(right is better). Mean and standard deviations (which are almost zero in some cases) are reported for each model. The results with both feature extractors are separated by the horizontal dashed line. The vertical, dashed lines represent the mean performance of the models using each feature extractor.



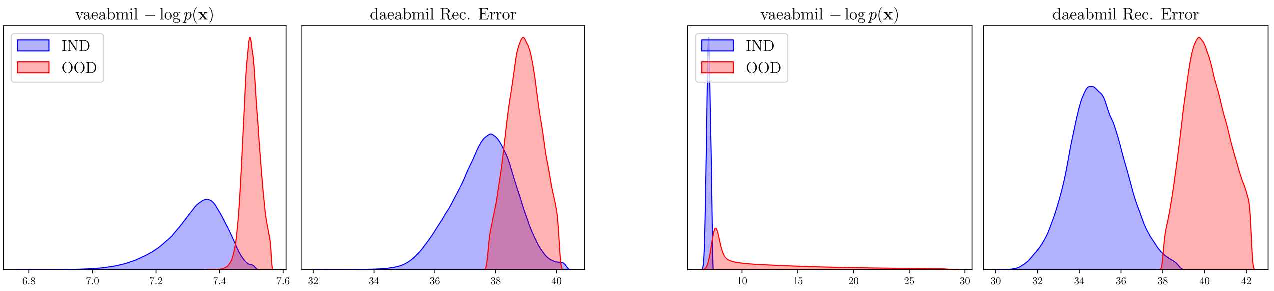
(a) (CAM16, PANDA)

(b) (PANDA, CAM16)

FIGURE 7. Approximated densities of the bag-level OOD scores produced by the models in the Far OOD detection experiments, using UNI features.

PANDA datasets. This highlights the benefits of the smooth, probabilistic latent space that the VAE in VAEABMIL produces. Also in Figure 6a, when using UNI features, the OOD detection performance of the rest of the models increases, due to the highly refined features that this foundation model produces. However, our models obtain the best result in OOD detection due to their explicit data-distribution modelling capability.

Figure 7a, shows the slide-level OOD score for all the models using UNI features. In this case, and although our proposals still perform better than the others, we observe



(a) (CAM16, PANDA)

(b) (PANDA, CAM16)

FIGURE 8. Approximated densities of the instance-level OOD scores produced by VAEABMIL and DAEABMIL in the Far OOD detection experiments using UNI features.

an also good performance by TRANSMIL and ABMIL, which also produce two different distributions for IND and OOD bags. Figure 8a shows the instance-level predicted OOD scores by our models, using UNI features.[4] The separation that our models produce is large enough to clearly distinguish between IND and OOD instances. This is coherent with the fact that the slides in this OOD detection problem contain different types of tissue.

2) (PANDA, CAM16)

Now, we use PANDA ad IND dataset and CAM16 as OOD dataset. Figure 6b shows the OOD detection results. The results are similar to those obtained in the previous experiment: we again observe that DAEABMIL performs on pair with the rest of the models. This supports our idea that the features produced by BT for PANDA and CAM16 are not discriminative enough to differentiate them through a deterministic autoencoder which produces a non-continuous latent space. VAEABMIL, however, obtains a perfect AUC score, highlighting the benefits of using a continuous, probabilistic latent space and modelling the likelihood of the data to detect out of distribution samples. When using UNI features, DAEABMIL and VAEABMIL are capable to detect all OOD bags correctly, outperforming the rest of the models.

Figure 7b depicts the bag-level OOD scores obtained by all the models, showing that thanks to UNI features, all the models separate the distributions of the IND and OOD sets, with VAEABMIL and DAEABMIL doing it perfectly. The good AUC results are supported by the correct instance-level OOD discrimination shown in Figure 8b, where in both cases we observe a instance-level separation between IND and OOD scores.

F. NEAR OOD DETECTION

To end the experimental section, we present the Near OOD detection problems where, as indicated in Section IV-B, the IND dataset and the OOD dataset share the main tissue type. The results shown in this section are supported by the statistical significance analysis performed in AppendixA.

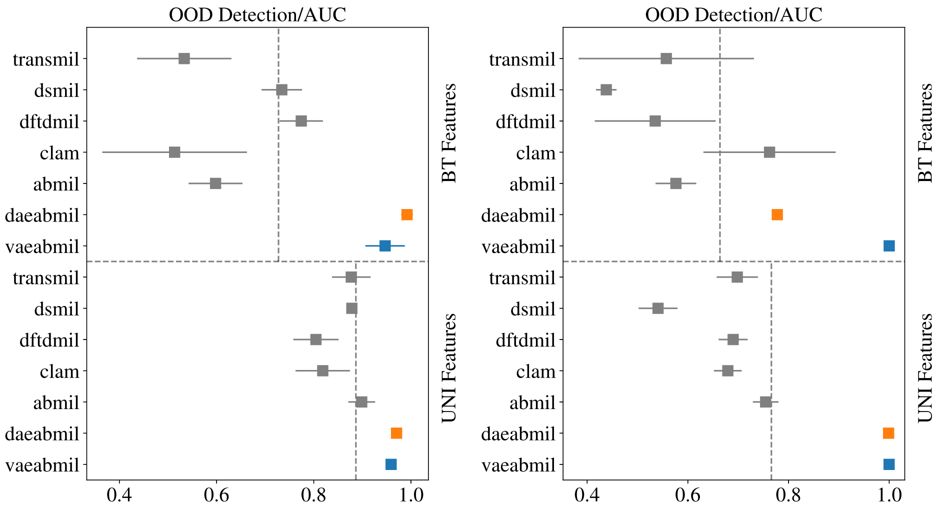
1) (CAM16, BCELL)

In this scenario, IND and OOD WSIs share the main tissue type but differ in their medical diagnosis. As described in Section IV-A, positive slides in CAM16 present cancer metastasis in lymph node sections, while BCELL WSIs have been diagnosed with diffuse large B-cell lymphoma. This poses, a priori, a more difficult OOD detection problem. Figure 9a shows, the OOD detection results. Observing this Figure, we highlight that:

- VAEABMIL and DAEABMIL excel at detecting OOD samples, obtaining an almost perfect AUC using any of the used features. This indicates that the autoencoders in both methods have learned to assign higher LOGPX and RECERR, respectively, to OOD samples than to IND ones.
- We observe considerably worse results for the rest of the models. When using BT features, the AUC is approximately 0.6 in some cases, indicating that the entropy of the predictions is the same for both IND and OOD data. This is an important problem with current SOTA MIL models, since their predictions are not well calibrated and can not detect OOD samples. This poses an important problem for their use in real diagnosis applications.
- When UNI features are used, the rest of the models show a strong improvement in the OOD detection, which correlates with the improvement in the classification AUC. We state that the foundation model UNI produces more discriminative features for the downstream tasks, separating OOD instances further away from IND data.

Figure 10a displays the WSI-level OOD score produced by each of the models. This figure reveals that the rest of the models assign very similar OOD scores to IND and OOD WSIs, which is a key drawback when using those models in a real world scenario like the one we are presenting. Our models, in contrast, produce separated distributions that may alert the pathologist when diagnosing a patient. Although TRANSMIL and ABMIL may seem to differentiate between IND and OOD distributions, the AUC metric in Figure 9a reveals that their OOD detection performance is still worse than VAEABMIL and DAEABMIL. Figure 11a, shows the instance-level OOD

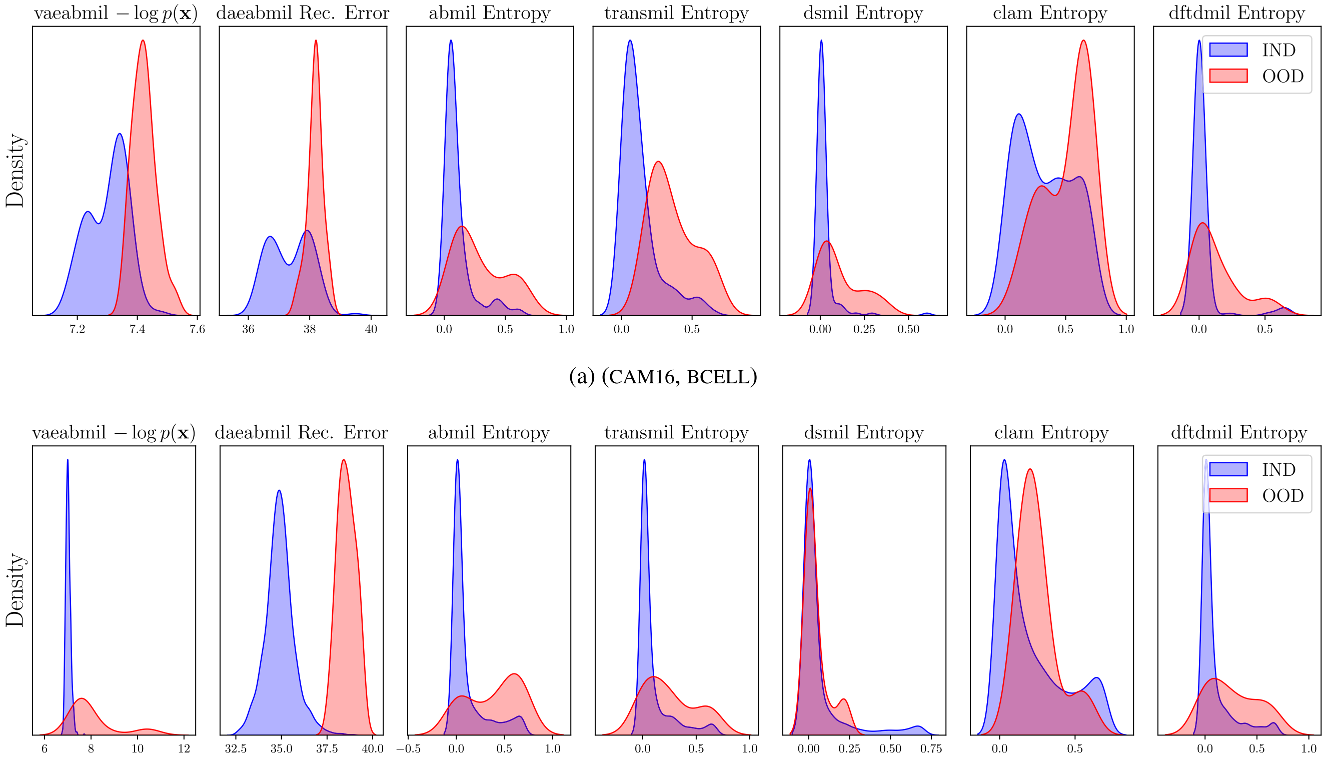
[4] Remark that the rest of the models can not predict instance-level OOD scores.



(a) (CAM16, BCELL)

(b) (PANDA, ARTIF)

FIGURE 9. Near OOD detection results. The presented metric is the AUC(right is better). Mean and standard deviations (which are almost zero in some cases) are reported for each model. The results with both feature extractors are separated by the horizontal dashed line. The vertical, dashed lines represent the mean performance of the models using each feature extractor.

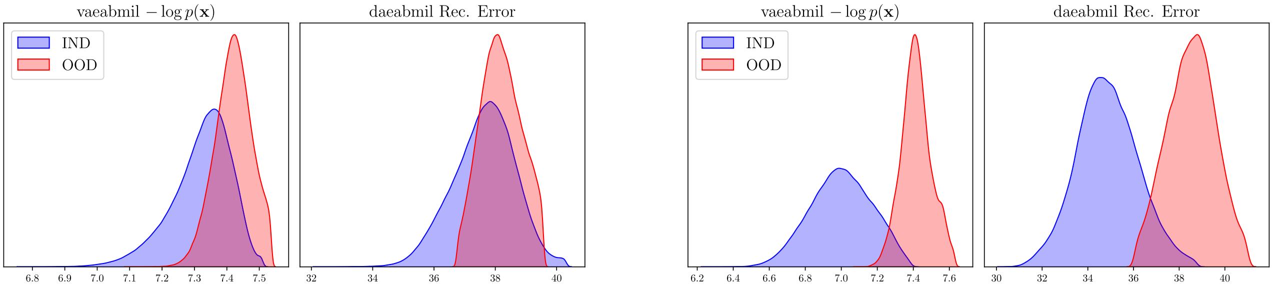


(a) (CAM16, BCELL)

(b) (PANDA ARTIF)

FIGURE 10. Approximated densities of the bag-level OOD scores produced by the models in the Near OOD detection experiments, using UNI features.

predictions of VAEABMIL and DAEABMIL, using UNI features. We observe that, even though there is overlapping between the estimated densities of the scores of IND and OOD instances, there is a shift in the mean of the distributions of IND and OOD instances, specially in VAEABMIL. Such distribution shift is the cause of the remarkable OOD detection capabilities of our models. Thanks to averaging the instance-level OOD scores, OOD bags are perfectly detected. Notice that the



(a) (CAM16, BCELL)

(b) (PANDA, ARTIF)

FIGURE 11. Approximated densities of the instance-level OOD scores produced by VAEABMIL and DAEABMIL in the Near OOD detection experiments using UNI features.

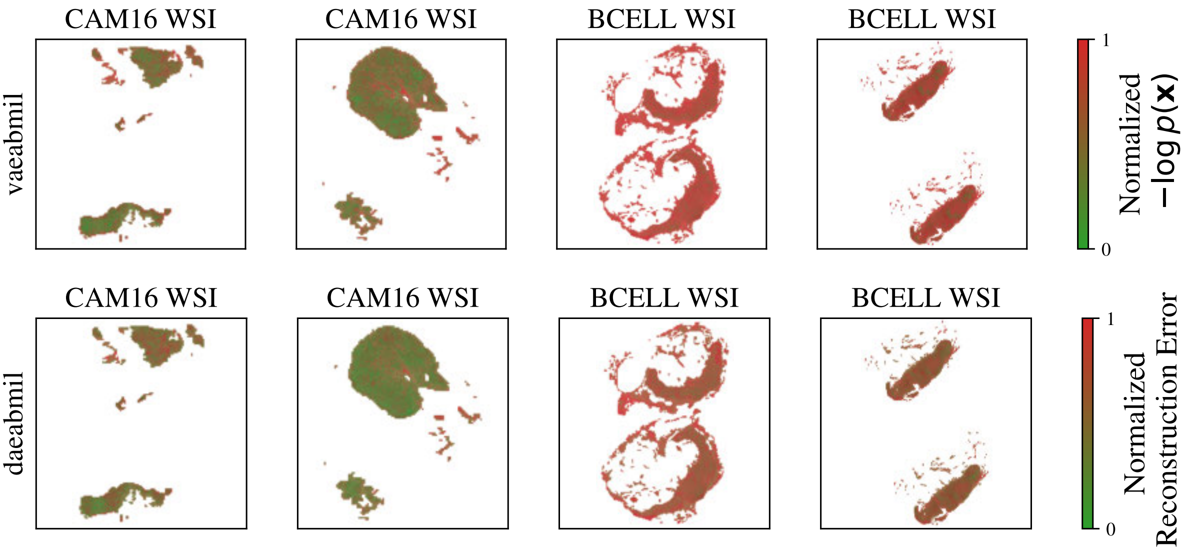
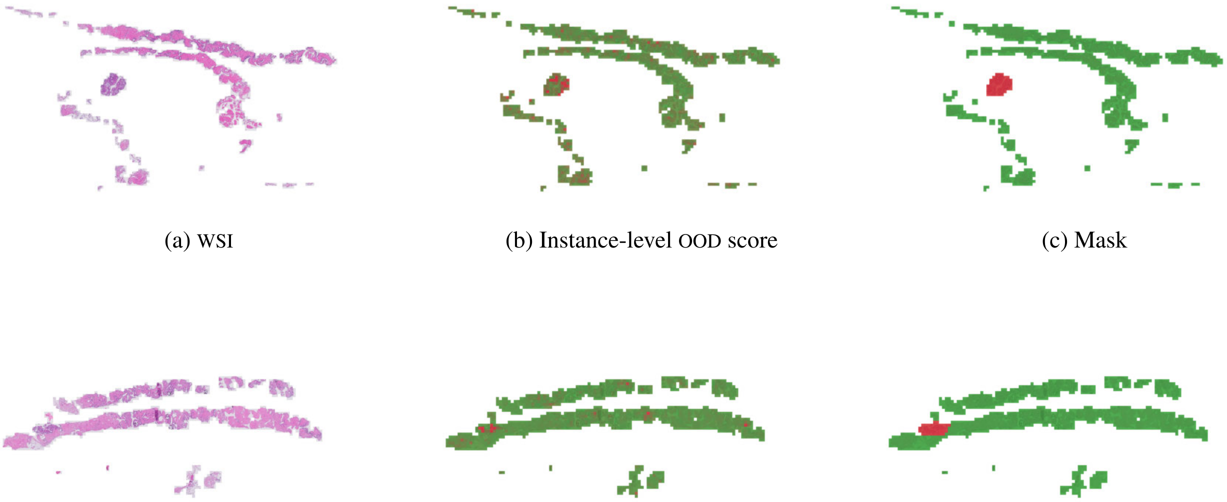

CAM16 WSI
CAM16 WSI
BCELL WSI
BCELL WSI
vaeabmil
daeabmil
1
0
Normalized $-\log p(\mathbf{x})$
1
0
Normalized Reconstruction Error

FIGURE 12. Top row: $-\log p(\mathbf{x})$ values obtained by VAEABMIL for each patch in both IND and OOD WSIs. Bottom Row: reconstruction error of each patch in both IND and OOD WSIs, obtained by DAEABMIL. In each rows, UNI features are used, and the predicted instance-level values are jointly normalized along the WSIs. VAEABMIL and DAEABMIL assign similar instance-level OOD scores in IND samples, being much higher in the OOD dataset (BCELL) than in the IND one (CAM16).



(a) WSI

(b) Instance-level OOD score

(c) Mask

(d) WSI

(e) Instance-level OOD score

(f) Mask

FIGURE 13. Visualization of two WSIs from the ARTIF dataset containing annotated artifacts. The corresponding instance-level OOD scores predicted by VAEABMIL and masks are shown. Each row corresponds to a different case. It is observed how VAEABMIL assigns higher OOD scores to the regions identified as artifacts in the mask.

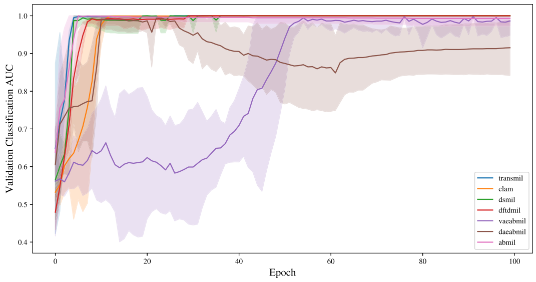


FIGURE 14. Validation AUC for all the trained models in the CAM16 dataset using features from UNI. Mean and 95% confidence intervals are shown per each model. The convergence of VAEABMIL is slower than that of the rest of the models. Also, DAEABMIL shows a performance decrease due to the double-objective optimization task.

instance-level OOD scores show which areas of the WSI are poorly reconstructed by the autoencoders and are, thus, more relevant to identify the slide as OOD. We show an example of this behaviour in Figure 12, where higher instance-level RECERR/LOGPX are obtained in BCELL (OOD) compared to the CAM16 (IND) patches.

2) (PANDA, ARTIF)

In the last experiment, we assess the OOD detection capabilities of our models in another real clinical scenario. Models trained in PANDA are evaluated by testing their ability to identify prostate slides containing pathologist-annotated artifacts. This represents a highly relevant case, as artifacts are commonly encountered in real-world WSIs. Figure 9b shows that our models, specially VAEABMIL, outperform the SOTA models in this task when using BT features. When using UNI features, the difference between our proposals and the SOTA models also becomes clear for DAEABMIL, which highlights the importance of using a foundation model as feature extractor for OOD detection tasks. This is also observed in the estimated densities of the bag-level OOD scores shown in Figure 10b. The conclusions are the same as the ones presented in Section IV-F1, showing consistency of our method. These results are clear indicator of the benefits of our proposal: our models present a novel approach that can perform the classification task on pair with the SOTA models while clearly outperforming them in the OOD detection task.

Figure 11b shows overlapping between IND and OOD instance-level distributions for this experiment. This is an expected behaviour since ARTIF contains prostate tissue as PANDA does. However, thanks to aggregating the scores in the whole bags, artifact-containing bags are correctly identified as OOD. Furthermore, in Figure 13 we leverage the instance-level OOD score provided by VAEABMIL to show that our proposal can be used to locate artifacts. This provides a visual tool for pathologists, adding to VAEABMIL high value for clinical use.

G. LIMITATIONS

Both proposed models exhibit one main limitation when compared to the other deep MIL models: our methods are harder to optimize than the rest. The reason for this is that, in both cases, the loss function to be optimized is composed of a classification-related term and two OOD detection-related terms. Thus, jointly optimizing all the terms compromises the effectiveness of the model, specially in the classification task, as we have observed in the results. This can also be observed in Figure 14, where we plot the classification AUC in the validation set during the optimization process in the CAM16 dataset using UNI features. We observe that VAEABMIL converges slower than the rest of the models. DAEABMIL, converges as fast as CLAM, but lowers its performance as the training process advances due to the need to also optimize for instance-level reconstruction task.

Nevertheless, even with this limitation, our proposals obtain comparable classification results and better OOD detection metrics, making them very useful in real-world scenarios.

V. CONCLUSION AND FUTURE WORK

While the apparition of OOD samples is very frequent in digital pathology, current MIL SOTA methods are not designed to reliably quantify whether a test bag belongs to the training data distribution. This limitation poses a great risk of incorrect predictions when unexpected tissues are encountered in real-world clinical settings. With this motivation, we propose a novel probabilistic deep MIL method with OOD capabilities. Our model, VAEABMIL, generalizes the well-known ABMIL using a VAE to model the data distribution, which gives the MIL method the ability to detect OOD samples by aggregating the marginal likelihood of the instances as an OOD score. Also, we have proposed a deterministic version, DAEABMIL, which leverages the reconstruction error as a deterministic OOD score. The main novelty of the proposed models is that they are defined and trained to perform two different tasks (bag-level classification and OOD detection) simultaneously, which none of the previous MIL methods is doing.

Extensive experimentation shows that VAEABMIL and DAEABMIL are competitive with the rest of the SOTA methods in the classification task. Furthermore, and very importantly for the design of CAD systems, they outperform current MIL methods at detecting OOD samples in both Near and Far OOD scenarios. The experiments also highlight the importance of using a foundation model as a feature extractor.

This work opens several promising directions for future research. One possibility is to extend the use of a VAE in combination with more complex MIL methods such as TRANSMIL or DTFDMIL. Another is to explore alternative generative models for learning the data distribution. Both approaches have the potential to enhance the OOD detection performance of MIL models.

APPENDIX A
STATISTICAL SIGNIFICANCE TEST

To assess whether differences in model performance in the OOD detection task are statistically significant, we employ

TABLE 1. T-test results comparing the OOD AUC in the (CAM16,PANDA) experiment using UNI features.

Model	OoD/auroc	t_stat	p_value	Significant
vaeabmil	0.9999 ± 0.0001	-	-	-
transmil	0.9690 ± 0.0059	8.9761	0.0122	True
clam	0.8300 ± 0.1985	1.4827	0.2764	False
dsmil	0.9438 ± 0.0224	4.3511	0.0490	True
dftdmil	0.9497 ± 0.0351	2.4722	0.1320	False
daeabmil	0.9923 ± 0.0001	352.6000	0.0000	True
abmil	0.9544 ± 0.0135	5.8111	0.0284	True

(a) Statistical comparison for VAEABMIL

Model	OoD/auroc	t_stat	p_value	Significant
daeabmil	0.9923 ± 0.0001	-	-	-
transmil	0.9690 ± 0.0059	6.7671	0.0211	True
clam	0.8300 ± 0.1985	1.4161	0.2924	False
dsmil	0.9438 ± 0.0224	3.7590	0.0640	False
dftdmil	0.9497 ± 0.0351	2.0995	0.1706	False
vaeabmil	0.9999 ± 0.0001	-352.6000	0.0000	True
abmil	0.9544 ± 0.0135	4.8494	0.0400	True

(b) Statistical comparison for DAEABMIL

TABLE 2. T-test results comparing the OOD AUC in the (CAM16,BCELL) experiment using UNI features.

Model	OoD/auroc	t_stat	p_value	Significant
vaeabmil	0.9592 ± 0.0030	-	-	-
transmil	0.8772 ± 0.0398	3.3333	0.0794	False
clam	0.8185 ± 0.0560	4.4629	0.0467	True
dsmil	0.8784 ± 0.0100	10.7940	0.0085	True
dftdmil	0.8045 ± 0.0466	5.7613	0.0288	True
daeabmil	0.9704 ± 0.0084	-1.7167	0.2282	False
abmil	0.8987 ± 0.0277	3.9603	0.0582	False

(a) Statistical comparison for VAEABMIL

Model	OoD/auroc	t_stat	p_value	Significant
daeabmil	0.9704 ± 0.0084	-	-	-
transmil	0.8772 ± 0.0398	4.8397	0.0401	True
clam	0.8185 ± 0.0560	4.4932	0.0461	True
dsmil	0.8784 ± 0.0100	38.1552	0.0007	True
dftdmil	0.8045 ± 0.0466	6.2819	0.0244	True
vaeabmil	0.9592 ± 0.0030	1.7167	0.2282	False
abmil	0.8987 ± 0.0277	4.0565	0.0557	False

(b) Statistical comparison for DAEABMIL

TABLE 3. T-test results comparing the OOD AUC in the (PANDA,CAM16) experiment using UNI features.

Model	OoD/auroc	t_stat	p_value	Significant
vaeabmil	1.0000 ± 0.0000	-	-	-
daeabmil	1.0000 ± 0.0000	2.0000	0.1835	False
dsmil	0.7926 ± 0.0306	11.7303	0.0072	True
clam	0.6966 ± 0.0516	10.1895	0.0095	True
dftdmil	0.9436 ± 0.0298	3.2799	0.0817	False
transmil	0.9106 ± 0.0223	6.9357	0.0202	True
abmil	0.9590 ± 0.0069	10.3725	0.0092	True

(a) Statistical comparison for VAEABMIL

Model	OoD/auroc	t_stat	p_value	Significant
daeabmil	1.0000 ± 0.0000	-	-	-
vaeabmil	1.0000 ± 0.0000	-2.0000	0.1835	False
dsmil	0.7926 ± 0.0306	11.7294	0.0072	True
clam	0.6966 ± 0.0516	10.1893	0.0095	True
dftdmil	0.9436 ± 0.0298	3.2798	0.0817	False
transmil	0.9106 ± 0.0223	6.9359	0.0202	True
abmil	0.9590 ± 0.0069	10.3694	0.0092	True

(b) Statistical comparison for DAEABMIL

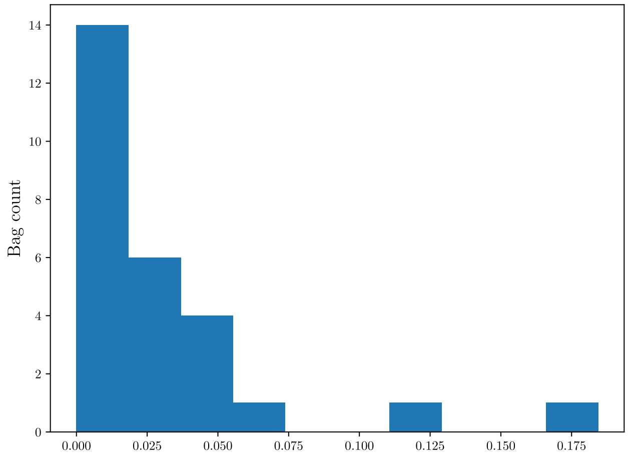


FIGURE 15. Histogram of the proportion of each WSI covered by an artifact in the ARTIF dataset. The proportion does not exceed 17.5%.

the paired t-test, a parametric test designed to compare two related samples [56]. In our case, since we executed $n = 3$ train/test partitions per model, we compare the OOD detection AUC of each model in each partition with the results of rest of the models in the same partition. The paired t-test evaluates whether the mean difference between these paired scores is significantly different from zero under the assumption that the differences are normally distributed. To achieve this, we compute the statistic $\text{t\_stat} = \bar{d} \cdot \sqrt{n}/s_d$, where $\bar{d}$ is the mean difference of the results in each of the n splits and s_d is the standard deviation of the differences. This test is ideal here because it accounts for the dependencies between the two sets of scores by considering that they were computed on the same data partitions.

Tables 1, 2, 3, and 4 show the results of comparing both VAEABMIL and DAEABMIL with current SOTA models across the different performed experiments using the UNI feature extractor. Using a level of significance of 0.05, the results show that:

- The differences between the results of VAEABMIL and DAEABMIL are not significant in any case. This highlights the fact that, when using UNI features, both models perform equally at detecting OOD samples.
- In the (PANDA, ARTIF) experiment, the differences between our proposals and the SOTA models are always significant, as shown in Table 4. The reason for this is that those methods do not model the data distribution, making post-hoc OOD scores worse in this scenario. Furthermore, the artifacts are, in proportion, much smaller than the main tissue, as shown in Figure 15.

TABLE 4. T-test results comparing the OOD AUC in the (PANDA,ARTIF) experiment using UNI features.

Model	OoD/auroc	t_stat	p_value	Significant
vaeabmil	0.9993 ± 0.0004	-	-	-
daeabmil	0.9988 ± 0.0001	2.2618	0.1521	False
dsmil	0.5408 ± 0.0386	20.3737	0.0024	True
clam	0.6794 ± 0.0276	20.3732	0.0024	True
dftdmil	0.6883 ± 0.0300	18.1359	0.0030	True
transmil	0.6982 ± 0.0410	12.8230	0.0060	True
abmil	0.7546 ± 0.0255	16.7655	0.0035	True

(a) Statistical comparison for VAEABMIL

Model	OoD/auroc	t_stat	p_value	Significant
daeabmil	0.9988 ± 0.0001	-	-	-
vaeabmil	0.9993 ± 0.0004	-2.2618	0.1521	False
dsmil	0.5408 ± 0.0386	20.5614	0.0024	True
clam	0.6794 ± 0.0276	20.0429	0.0025	True
dftdmil	0.6883 ± 0.0300	17.9140	0.0031	True
transmil	0.6982 ± 0.0410	12.6833	0.0062	True
abmil	0.7546 ± 0.0255	16.5333	0.0036	True

(b) Statistical comparison for DAEABMIL

TABLE 5. Tables with the OOD detection results using multiple OOD scores. MLS stands for Maximum Logit score. In the scores defined for VAEABMIL and DAEABMIL, MAX and MEAN indicate the Maximum aggregator and the Mean aggregator, respectively.

Model	OoD/Entropy/auc	OoD/MLS/auc	OoD/LOGPXMAX/auc	OoD/LOGPXMEAN/auc	OoD/RECERRMAX/auc	OoD/RECERRMEAN/auc
abmil	0.954 ± 0.013	0.935 ± 0.031	-	-	-	-
clam	0.830 ± 0.199	0.826 ± 0.203	-	-	-	-
daeabmil	0.963 ± 0.011	0.968 ± 0.007	-	-	0.457 ± 0.031	0.992 ± 0.000
dftdmil	0.950 ± 0.035	0.961 ± 0.015	-	-	-	-
dsmil	0.944 ± 0.022	0.923 ± 0.019	-	-	-	-
transmil	0.969 ± 0.006	0.960 ± 0.013	-	-	-	-
vaeabmil	0.979 ± 0.007	0.970 ± 0.011	0.680 ± 0.109	1.000 ± 0.000	-	-

(a) OOD detection results for the different OOD scores in the (CAM16, PANDA) experiment.

Model	OoD/Entropy/auc	OoD/MLS/auc	OoD/LOGPXMAX/auc	OoD/LOGPXMEAN/auc	OoD/RECERRMAX/auc	OoD/RECERRMEAN/auc
abmil	0.899 ± 0.028	0.867 ± 0.046	-	-	-	-
clam	0.726 ± 0.071	0.722 ± 0.070	-	-	-	-
daeabmil	0.891 ± 0.027	0.916 ± 0.027	-	-	0.848 ± 0.006	0.970 ± 0.008
dftdmil	0.803 ± 0.049	0.790 ± 0.053	-	-	-	-
dsmil	0.878 ± 0.010	0.864 ± 0.003	-	-	-	-
transmil	0.877 ± 0.040	0.836 ± 0.072	-	-	-	-
vaeabmil	0.882 ± 0.035	0.877 ± 0.022	0.828 ± 0.027	0.959 ± 0.003	-	-

(b) OOD detection results for the different OOD scores in the (CAM16, BCELL) experiment.

Model	OoD/Entropy/auc	OoD/MLS/auc	OoD/LOGPXMAX/auc	OoD/LOGPXMEAN/auc	OoD/RECERRMAX/auc	OoD/RECERRMEAN/auc
abmil	0.959 ± 0.007	0.956 ± 0.005	-	-	-	-
clam	0.697 ± 0.052	0.654 ± 0.045	-	-	-	-
daeabmil	0.333 ± 0.156	0.334 ± 0.156	-	-	0.999 ± 0.000	1.000 ± 0.000
dftdmil	0.944 ± 0.030	0.949 ± 0.025	-	-	-	-
dsmil	0.793 ± 0.031	0.833 ± 0.032	-	-	-	-
transmil	0.911 ± 0.022	0.888 ± 0.033	-	-	-	-
vaeabmil	0.965 ± 0.016	0.982 ± 0.013	1.000 ± 0.000	1.000 ± 0.000	-	-

(c) OOD detection results for the different OOD scores in the (PANDA, CAM16) experiment.

Model	OoD/Entropy/auc	OoD/MLS/auc	OoD/LOGPXMAX/auc	OoD/LOGPXMEAN/auc	OoD/RECERRMAX/auc	OoD/RECERRMEAN/auc
abmil	0.755 ± 0.025	0.771 ± 0.027	-	-	-	-
clam	0.679 ± 0.028	0.646 ± 0.029	-	-	-	-
daeabmil	0.327 ± 0.076	0.344 ± 0.088	-	-	0.998 ± 0.001	0.999 ± 0.000
dftdmil	0.689 ± 0.029	0.704 ± 0.036	-	-	-	-
dsmil	0.541 ± 0.039	0.566 ± 0.037	-	-	-	-
transmil	0.698 ± 0.041	0.688 ± 0.049	-	-	-	-
vaeabmil	0.738 ± 0.026	0.771 ± 0.036	1.000 ± 0.000	0.999 ± 0.000	-	-

(d) OOD detection results for the different OOD scores in the (PANDA, ARTIF) experiment.

- In the (CAM16, PANDA), (CAM16, BCELL) and (PANDA, CAM16) experiments, some of the SOTA models obtain non-significant differences according to the paired t-test. However, we observe that the standard deviation of the OOD/auroc in our models is much smaller than in the rest of the models. Hence, we state that if the high variance was maintained when increasing the number of executions, the results would change to obtain significant differences between our proposals and the SOTA MIL models.

APPENDIX B
RESULTS WITH ALL THE OOD SCORES

To provide a comprehensive analysis of the performance of the different OOD scores across the used models, we present the complete results in Table 5. The results show that, for

ABMIL, TRANSMIL, and CLAM, the Entropy score obtained the highest OOD detection results in all cases. In DTFDMIL, the MLS achieves the best results except in the (CAM16, BCELL) scenario. In DSMIL, the Entropy is a better OOD detector when the IND dataset is CAM16, and the MLS performs better when the IND dataset is PANDA.

Regarding the proposed models, we observe that using the MEAN aggregator yields better results than the MAX aggregator in all cases except for the (PANDA, ARTIF) experiment. The difference between the aggregator is, however, negligible in that case. In other cases, such as when considering CAM16 as the IND dataset, the MAX aggregator struggles to detect OOD samples while applying the mean aggregator provides much better results.

In summary, when using the MEAN aggregator for LOGPX in VAEABMIL and for RECERR in DAEABMIL, the proposed models obtain the best performance in the OOD detection task.

ACKNOWLEDGMENT

The authors would like to thank Prof. Geert Litjens from Radboud University Medical Center for providing the BCELL dataset.

REFERENCES

[1] M. Gadermayr and M. Tschuchnig, "Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential," *Computerized Med. Imag. Graph.*, vol. 112, Mar. 2024, Art. no. 102337.

[2] M. Waqas, S. U. Ahmed, M. A. Tahir, J. Wu, and R. Qureshi, "Exploring multiple instance learning (MIL): A brief survey," *Expert Syst. Appl.*, vol. 250, Sep. 2024, Art. no. 123893.

[3] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 10, 1997, pp. 1–7.

[4] J. Foulds and E. Frank, "A review of multi-instance learning assumptions," *Knowl. Eng. Rev.*, vol. 25, no. 1, pp. 1–25, Mar. 2010.

[5] F. Herrera, S. Ventura, R. Bello, C. Cornelis, A. Zafra, D. Sánchez-Tarragó, and S. Vluymans, *Multiple Instance Learning*. Cham, Switzerland: Springer, 2016.

[6] E. Raff and J. Holt, "Reproducibility in multiple instance learning: A case for algorithmic unit tests," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 1–15.

[7] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Med.*, vol. 25, no. 8, pp. 1301–1309, Aug. 2019.

[8] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artif. Intell.*, vol. 201, pp. 81–105, Aug. 2013.

[9] O. Fourkioti, M. D. Vries, and C. Bakal, "CAMIL: Context-aware multiple instance learning for cancer detection and subtyping in whole slide images," in *Proc. ICLR*, 2024, pp. 1–16.

[10] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognit.*, vol. 74, pp. 15–24, Feb. 2018.

[11] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. ICML*, Jan. 2018, pp. 2127–2136.

[12] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, and X. Ji, "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 2136–2147.

[13] Y. Zhao, Z. Lin, K. Sun, Y. Zhang, J. Huang, L. Wang, and J. Yao, "SETMIL: Spatial encoding transformer-based multiple instance learning for pathological image analysis," in *Proc. MICCAI*, Jan. 2022, pp. 66–76.

[14] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14313–14323.

[15] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S. E. Coupland, and Y. Zheng, "DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18780–18790.

[16] F. M. Castro-Macías, P. Morales Alvarez, Y. Wu, R. Molina, and A. Katsaggelos, "SM: Enhanced localization in Multiple Instance Learning for medical imaging classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 77494–77524.

[17] Y. Wu, F. M. Castro-Macías, P. Morales-Álvarez, R. Molina, and A. K. Katsaggelos, "Smooth attention for deep multiple instance learning: Application to ct intracranial hemorrhage detection," in *Proc. MICCAI*, 2023, pp. 327–337.

[18] I. Irmakci, R. Nateghi, R. Zhou, M. Vescovo, M. Saft, A. E. Ross, X. J. Yang, L. A. D. Cooper, and J. A. Goldstein, "Tissue contamination challenges the credibility of machine learning models in real world digital pathology," *Mod. Pathol.*, vol. 37, no. 3, Mar. 2024, Art. no. 100422.

[19] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 427–436.

[20] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Deep learning for medical anomaly detection—A survey," *ACM Comput. Surv.*, vol. 54, no. 7, pp. 1–37, 2021.

[21] N. Kanwal, M. López-Pérez, U. Kiraz, T. C. M. Zuiverloon, R. Molina, and K. Engan, "Are you sure it's an artifact? Artifact detection and uncertainty quantification in histological images," *Computerized Med. Imag. Graph.*, vol. 112, Mar. 2024, Art. no. 102321.

[22] B. Schömig-Markiefka, A. Pryalukhin, W. Hulla, A. Bychkov, J. Fukuoka, A. Madabhushi, V. Achter, L. Nieroda, R. Büettner, A. Quaas, and Y. Tolkach, "Quality control stress test for deep learning-based diagnostic model in digital pathology," *Mod. Pathol.*, vol. 34, no. 12, pp. 2098–2108, Dec. 2021.

[23] J. Linmans, S. Elfwing, J. van der Laak, and G. Litjens, "Predictive uncertainty estimation for out-of-distribution detection in digital pathology," *Med. Image Anal.*, vol. 83, Jan. 2023, Art. no. 102655.

[24] A. Guha Roy et al., "Does your dermatology classifier know what it doesn't know? Detecting the long-tail of unseen conditions," *Med. Image Anal.*, vol. 75, Jan. 2022, Art. no. 102274.

[25] M. S. Graham, P.-D. Tudosiu, P. Wright, W. H. L. Pinaya, U. Jean-Marie, Y. H. Mah, J. T. Teo, R. Jager, D. Werring, P. Nachev, S. Ourselin, and M. J. Cardoso, "Transformer-based out-of-distribution detection for clinically safe segmentation," in *Proc. MIDL*, 2022, pp. 457–476.

[26] S. Basart, M. Mantas, M. Mostajabi, J. Steinhardt, and D. Song, "Scaling out-of-distribution detection for real-world settings," in *Proc. ICML*, 2022, pp. 1–14.

[27] G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard, "Multiple-instance learning for medical image and video analysis," *IEEE Rev. Biomed. Eng.*, vol. 10, pp. 213–234, 2017.

[28] D. Zimmerer, P. M. Full, F. Isense, and P. Jager, "MOOD 2020: A public benchmark for out-of-distribution detection and localization on medical images," *IEEE Trans. Med. Imag.*, vol. 41, no. 10, pp. 2728–2738, Oct. 2022.

[29] B. Kompa, J. Snoek, and B. A. L. Beam, "Second opinion needed: Communicating uncertainty in medical machine learning," *NPJ Digit. Med.*, vol. 4, p. 4, Jan. 2021.

[30] T. Araujo, G. Aresta, U. Schmidt-Erfurth, and H. Bogunović, "Few-shot out-of-distribution detection for automated screening in retinal OCT images using deep learning," *Sci. Rep.*, vol. 13, no. 1, p. 16231, Sep. 2023.

[31] L. Yan, F. Wang, L. Leng, and A. B. J. Teoh, "Toward comprehensive and effective palmprint reconstruction attack," *Pattern Recognit.*, vol. 155, Nov. 2024, Art. no. 110655.

[32] J. Linmans, J. van der Laak, and G. Litjens, "Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks," in *Proc. Conf. MIDL*, vol. 121, Jul. 2020, pp. 465–478.

[33] P. Seeböck, J. I. Orlando, T. Schlegl, S. M. Waldstein, H. Bogunovic, S. Klimscha, G. Langs, and U. Schmidt-Erfurth, "Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal OCT," *IEEE Trans. Med. Imag.*, vol. 39, no. 1, pp. 87–98, Jan. 2020.

[34] J. Linmans, G. Raya, J. van der Laak, and G. Litjens, "Diffusion models for out-of-distribution detection in digital pathology," *Med. Image Anal.*, vol. 93, Apr. 2024, Art. no. 103088.

[35] M. Pocevičiūtė, Y. Ding, R. Bromée, and G. Eilertsen, "Out-of-distribution detection in digital pathology: Do foundation models bring the end to reconstruction-based approaches?" *Comput. Biol. Med.*, vol. 184, Jan. 2025, Art. no. 109327.

[36] L. Nie, F. Jiao, W. Wang, Y. Wang, and Q. Tian, "Conversational image search," *IEEE Trans. Image Process.*, vol. 30, pp. 7732–7743, 2021.

[37] L. Nie, W. Wang, R. Hong, M. Wang, and Q. Tian, "Multimodal dialog system: Generating responses via adaptive decoders," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1098–1106.

[38] X. Ran, M. Xu, L. Mei, Q. Xu, and Q. Liu, "Detecting out-of-distribution samples via variational auto-encoder with reliable uncertainty estimation," *Neural Netw.*, vol. 145, pp. 199–208, Jan. 2022.

[39] Y. Wu, P. Besson, E. A. Azcona, S. Kathleen Bandt, T. B. Parrish, and A. K. Katsaggelos, "Reconstruction of resting state FMRI using LSTM variational auto-encoder on subcortical surface to detect epilepsy," in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Mar. 2022, pp. 1–5.

[40] E. Daxberger and J. M. Hernández-Lobato, "Bayesian variational autoencoders for unsupervised out-of-distribution detection," 2020, *arXiv:1912.05651*.

[41] Q. Zhou, S. Wang, X. Zhang, and Y.-D. Zhang, "WVALE: Weak variational autoencoder for localisation and enhancement of COVID-19 lung infections," *Comput. Methods Programs Biomed.*, vol. 221, Jun. 2022, Art. no. 106883.

[42] H. Liu, Z. Zhao, X. Chen, R. Yu, and Q. She, "Using the VQ-VAE to improve the recognition of abnormalities in short-duration 12-lead electrocardiogram records," *Comput. Methods Programs Biomed.*, vol. 196, Nov. 2020, Art. no. 105639.

[43] W. Zhang, X. Zhang, and M. L. Zhang, "Multi-instance causal representation learning for instance label prediction and out-of-distribution generalization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 34940–34953.

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

[45] C. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[46] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-VAE," 2018, *arXiv:1804.03599*.

[47] B. E. Bejnordi, M. Veta, P. J. V. Diest, B. V. Ginneken, N. Karssemeijer, G. Litjens, and J. A. W. M. van der Laak, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, no. 22, pp. 2199–2210, 2017.

[48] W. Bulten et al., "Artificial intelligence for diagnosis and Gleason grading of prostate cancer: The PANDA challenge," *Nature Med.*, vol. 28, no. 1, pp. 154–163, Jan. 2022.

[49] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. ICML*, 2021, pp. 12310–12320.

[50] M. Kang, H. Song, S. Park, D. Yoo, and S. Pereira, "Benchmarking self-supervised learning on diverse pathology datasets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3344–3354.

[51] R. J. Chen, T. Ding, M. Y. Lu, D. F. K. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, M. Williams, L. Oldenburg, L. L. Weishaupt, J. J. Wang, A. Vaidya, L. P. Le, G. Gerber, S. Sahai, W. Williams, and F. Mahmood, "Towards a general-purpose foundation model for computational pathology," *Nature Med.*, vol. 30, no. 3, pp. 850–862, Mar. 2024.

[52] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomed. Eng.*, vol. 5, no. 6, pp. 555–570, Mar. 2021.

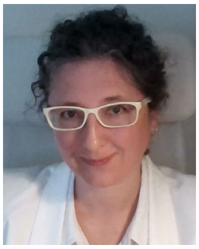
[53] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997.

[54] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017, *arXiv:1412.6980*.

[56] A. Ross and V. L. Willson, "Paired samples t-test," in *Basic and Advanced Statistical Tests: Writing Results Sections and Creating Tables and Figures*. Rotterdam, The Rotterdam: SensePublishers, 2017, pp. 17–19.

FRANCISCO JAVIER SÁEZ-MALDONADO received the dual B.Sc. degree in mathematics and computer science from Universidad de Granada, in 2021, and the M.S. degree in data science from the Autonomous University of Madrid, in 2023. He is currently pursuing the Ph.D. degree with Universidad de Granada under the supervision of Prof. Molina and Prof. Morales-Lvarez. His research interests include Bayesian modeling and uncertainty estimation, with a particular focus on likelihood methods for out-of-distribution detection and Gaussian processes, and their applications to medical imaging problems.



LUZ GARCÍA received the M.Sc. degree in telecommunication engineering from the Polytechnic University of Madrid, Madrid, Spain, in 2000, and the Ph.D. degree in telecommunication engineering from Universidad de Granada, Granada, Spain, in 2008. After, she was a Support Engineer in communication networks with Ericsson-Spain, Madrid, from 2000 to 2004. She joined a European Research Project with Universidad de Granada. She was an Assistant Professor with the Department of Signal Theory, Telematics, and Communications, Universidad de Granada, where she has been a Senior Lecturer, since 2019. Her research interests include signal processing, pattern recognition, and machine learning in the fields of biometrics, distributed acoustic sensing, and seismology.

LEE A. D. COOPER received the Ph.D. degree in electrical and computer engineering from The Ohio State University, in 2009. He joined the Biomedical Informatics Faculty, Emory University, in 2012, where he was jointly appointed with the Department of Biomedical Engineering, Georgia Institute of Technology. He joined the Department of Pathology, Northwestern University, in 2019, as an Associate Professor, and the Director of Computational Pathology.

JEFFERY A. GOLDSTEIN received the M.D. and Ph.D. degrees from the University of Chicago, where he was struck by the lack of evidence-based treatment in obstetrics and the significant burden that premature birth and infant health complications place on families and caregivers. He completed a Residency in anatomic pathology from Vanderbilt University, and a Fellowship in pediatric pathology from Lurie Children's Hospital and Northwestern University. During his residency and fellowship, he conducted both clinical and research work in maternal-child health, incorporating bioimaging, and informatics. He is an early-stage Investigator utilizing bioimaging and informatics techniques to enhance the diagnosis and treatment of maternal-child health issues. He is an attending Physician with Northwestern Memorial Hospital, where he has clinical and teaching responsibilities in perinatal and autopsy pathology. His research primarily focuses on the examination of microscopic slides from placentas. These experiences have established him as a content expert in placental pathology with a strong foundation in computational methods. His proposed project aims to build on these skills and further establish him as an independent researcher in the field.

RAFAEL MOLINA (Life Senior Member, IEEE) received the M.Sc. degree in mathematics (statistics) and the Ph.D. degree in optimal design in linear models from Universidad de Granada, Granada, Spain, in 1979 and 1983, respectively. He was the Dean of the School of Computer Engineering, Universidad de Granada, from 1992 to 2002, where he became a Professor of computer science and artificial intelligence, in 2000. He was the Head of the Department of Computer Science and Artificial Intelligence, Universidad de Granada, from 2005 to 2007. He has co-authored an article that received the Runner-Up Prize from the Reception for Early Stage Researchers at the House of Commons, in 2007, the Best Student Paper from the IEEE International Conference on Image Processing, in 2007, the ISPA Best Paper, in 2009, and the EUSIPCO 2013 Best Student Paper. His research interests include Bayesian modeling and inference in image restoration (applications to astronomy and medicine), super-resolution of images and video, blind deconvolution, computational photography, source recovery in medicine, compressive sensing, low-rank matrix decomposition, active learning, fusion, supervised learning, and crowdsourcing. He has served as an Associate Editor for *Applied Signal Processing*, from 2005 to 2007, and IEEE Transactions on Image Processing, from 2010 to 2014. Since 2011, he has been serving as an Area Editor for *Digital Signal Processing*.

AGGELOS K. KATSAGGELOS (Life Fellow, IEEE) received the Diploma degree in electrical and mechanical engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 1979, and the M.S. and Ph.D. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, GA, USA, in 1981 and 1985, respectively. In 1985, he joined the Department of Electrical Engineering and Computer Science, Northwestern University, where he is currently a Professor Holder with the Joseph Cummings Chair. Previously, he was the Holder of the Ameritech Chair of Information Technology and the AT&T Chair. He is also a member of the Academic Staff, NorthShore University Health System, an affiliated Faculty Member of the Department of Linguistics, and he has an appointment with the Argonne National Laboratory. He has authored or co-authored extensively in the areas of multimedia signal processing and communications, computational imaging, and machine learning, including more than 250 journal articles, 600 conference papers, and 40 book chapters, and he is the holder of 30 international patents. He is the co-author of *Rate-Distortion Based Video Compression* (Kluwer), in 1997, *Super-Resolution for Images and Video* (Claypool), in 2007, *Joint Source-Channel Video Transmission* (Claypool), in 2007, and *Machine Learning Refined* (Cambridge University Press), in 2016. He has supervised 57 Ph.D. theses. Among his many professional activities, he was a BOG Member of the IEEE Signal Processing Society, from 1999 to 2001, a member of the Publication Board of Proceedings of the IEEE, from 2003 to 2007, and a member of the Award Board of the IEEE Signal Processing Society. He was a fellow of SPIE, in 2009, EURASIP, in 2017, and OSA, in 2018. He was a recipient of the IEEE Third Millennium Medal, in 2000, the IEEE Signal Processing Society Meritorious Service Award, in 2001, the IEEE Signal Processing Society Technical Achievement Award, in 2010, the IEEE Signal Processing Society Best Paper Award, in 2001, the IEEE ICME Paper Award, in 2006, the IEEE ICIP Paper Award, in 2007, the ISPA Paper Award, in 2009, and the EUSIPCO Paper Award, in 2013. He was a Distinguished Lecturer of the IEEE Signal Processing Society, from 2007 to 2008. He was the Editor-in-Chief of *IEEE Signal Processing Magazine*, from 1997 to 2002.