Computer Methods and Programs in Biomedicine 224 (2022) 107012

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Constrained multiple instance learning for ulcerative colitis prediction using histological images

Rocío del Amor[a,*], Pablo Meseguer[a], Tommaso Lorenzo Parigi[b,c], Vincenzo Villanacci[d], Adrián Colomer[a], Laëtitia Launet[a], Alina Bazarova[g], Gian Eugenio Tontini[h], Raf Bisschops[i], Gert de Hertogh[i], Jose G. Ferraz[j], Martin Götz[k], Xianyong Gui[y], Bu'Hussain Hayee[m], Mark Lazarev[n], Remo Panaccione[j], Adolfo Parra-Blanco[o], Pradeep Bhandari[l], Luca Pastorelli[p], Timo Rath[q], Elin Synnøve Røyset[r], Michael Vieth[s,x], Davide Zardo[t], Enrico Grisan[v,w], Subrata Ghosh[u,g], Marietta Iacucci[b,e,f], Valery Naranjo[a]

[a] Instituto de Investigación e Innovación en Bioingeniería, Universitat Politécnica de Valéncia, Valencia, Spain
[b] Department of Biomedical Sciences, Humanitas University, Milan, Italy
[c] University of Birmingham, Immunology and Immunotherapy, Birmingham, United Kingdom
[d] Institute of Pathology, ASST Spedali Civili, University of Brescia, Brescia, Italy
[e] National Institute for Health Research (NIHR) Biomedical Research Centre, Birmingham, United Kingdom
[f] Department of Gastroenterology, University Hospitals Birmingham NHS Trust, Birmingham, United Kingdom
[g] Institute for Biological Physics, University of Cologne, Cologne, Germany
[h] Fondazione IRCCS Ca'Granda Ospedale Maggiore Policlinico, Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy
[i] Division of Gastroenterology, University Hospitals Leuven, Leuven, Belgium
[j] Division of Gastroenterology, University of Calgary Cumming School of Medicine, Calgary, Canada
[k] Division of Gastroenterology, Klinikum, Böblingen, Germany
[l] Division of Gastroenterology, Queen Alexandra Hospital, Portsmouth, United Kingdom
[m] Division of Gastroenterology, Kings College London, London, United Kingdom
[n] Division of Gastroenterology, Johns Hopkins Hospital, Baltimore, United States
[o] Division of Gastroenterology, University of Nottingham, Nottingham, United Kingdom
[p] Liver and Gastroenterology Unit, Universita' degli Studi di Milano, ASST Santi Paolo E Carlo, University Hospital San Paolo, Milan, Italy
[q] Division of Gastroenterology, University of Erlangen, Erlangen, Germany
[r] Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway
[s] Klinikum Bayreuth, Bayreuth, Germany
[t] Department of Pathology, San Bortolo Hospital, Vicenza, Italy
[u] APC Microbiome Ireland, College of Medicine and Health, Cork, Ireland
[v] Department of Information Engineering, Padova, Italy
[w] School of Engineering, London South Bank University, London, UK
[x] Institute of Pathology, Friedrich-Alexander-University Erlangen-Nuremberg, Nuremberg, Germany
[y] Department of Laboratory Medicine and Pathology, University of Washington, Seattle, USA

ARTICLE INFO

Article history:
Received 6 May 2022
Revised 5 July 2022
Accepted 6 July 2022

Keywords:
Histologic remission
Location constraints
Neutrophils
Attention-embedding weights
Ulcerative colitis

ABSTRACT

Background and Objective: Ulcerative colitis (UC) is an inflammatory bowel disease (IBD) affecting the colon and the rectum characterized by a remitting-relapsing course. To detect mucosal inflammation associated with UC, histology is considered the most stringent criteria. In turn, histologic remission (HR) correlates with improved clinical outcomes and has been recently recognized as a desirable treatment target. The leading biomarker for assessing histologic remission is the presence or absence of neutrophils. Therefore, the finding of this cell in specific colon structures indicates that the patient has UC activity. However, no previous studies based on deep learning have been developed to identify UC based on neutrophils detection using whole-slide images (WSI).

Methods: The methodological core of this work is a novel multiple instance learning (MIL) framework with location constraints able to determine the presence of UC activity using WSI. In particular, we put forward an effective way to introduce constraints about positive instances to effectively explore additional weakly supervised information that is easy to obtain and enjoy a significant boost to the learning process. In addition, we propose a new weighted embedding to enlarge the relevance of the positive instances.

Results: Extensive experiments on a multi-center dataset of colon and rectum WSIs, PICASSO-MIL, demonstrate that using the location information we can improve considerably the results at WSI-level. In comparison with prior MIL settings, our method allows for 10% improvements in bag-level accuracy.

Conclusion: Our model, which introduces a new form of constraints, surpass the results achieved from current state-of-the-art methods that focus on the MIL paradigm. Our method can be applied to other histological concerns where the morphological features determining a positive WSI are tiny and similar to others in the image.

https://doi.org/10.1016/j.cmpb.2022.107012

Table 1
PICaSSO Histologic Remission Index (PHRI) to predict histological remission.

Histologic finding	Score
Neutrophil infiltration in lamina propria	
Absent (No)	0
Present (Yes)	1
Neutrophil infiltration in epithelium	
Absent (No)	0
Present (Yes)	
- Surface epithelium	1
- Cryptal epithelium	1
- Crypt abscess	1
Total Score = sum of all above (maximum 4)	

1. Introduction

Ulcerative colitis (UC) is a chronic inflammatory bowel disease (IBD) affecting the colon and the rectum with a propensity to arise in adolescents and young adults. The incidence of UC has been increasing globally [1] and currently ranges from 4 to 20 per 100,000 in North America and Europe [2].

The treatment of UC aims to extinguish bowel inflammation and prevent complications. Histological assessment plays a critical role in determining inflammatory activity. In this vein, histologic remission (HR) (also referred to as histologic healing, HH) is emerging as the most rigorous target of treatment and is associated with favorable clinical outcomes [3–6]. However, incorporating histology into clinical practice remains challenging. This is due to: (1) the lack of a universal definition of HR that varies depending on the histological score/index applied, (2) the complexity of most scores and (3) the high inter-observer variability between pathologists [4,7–9].

Over the past decades, more than 30 histological scores have been developed, although their adoption in clinical practice remains modest [10,11]. Similarly, different definitions and criteria of HR have been proposed, ranging from 'elimination of mucosal ulceration/erosion' to 'complete histological normalization'. Almost all investigators now agree that the absence of neutrophilic infiltration ('neutrophil-free' mucosa) is the key to define HR [11–14]. Indeed, this has been endorsed by two independent expert panels [14,15]. Recently, our medical team developed a simplified histological score, PICASSO Histological Remission Index or PHRI, see Table 1 [16].

The primary aim of PHRI was to create a simple 'neutrophil only' histologic evaluation that predicted specified clinical outcomes. The structures of the biopsy where to evaluate the presence or absence of neutrophils and predict histological remission are: (a) lamina propia, (b) surface epithelium, (c) cryptal epithelium and (d) cryptal lumen, see Fig. 1.

The computer-aided diagnosis systems (CADs) based on artificial intelligence (AI) aim to support pathologists in the daily analysis of histological biopsies, reducing both the workload and the inconsistency generated. Their final goal is to produce a reliable and reproducible real-time assessment of disease activity. With the emergence of digital pathology, the digitization of histological tissue sections into whole-slide images (WSIs) has been standardized, leading to the application of computer vision methods. Additionally, previous research showed the applicability of computer vision methods based on deep-learning approaches using WSIs for cancer detection, inflammatory prediction, etc. Regarding the detection of UC activity based on deep learning techniques, available research has focused on the analysis of endoscopic images [17–21], but so far, only one study has approached the analysis of WSIs [22]. In [22], the authors used a deep learning algorithm to quantify the density of eosinophils in sigmoid colon biopsies from consecutive UC patients with histologically active disease. The algorithm was applied to sigmoid and colon biopsies from a cross-sectional cohort of 88 UC patients with histologically active disease as measured by the Geboes score and Robarts histopathology index (RHI). However, this study does not differentiate between remission and active WSI.

To the best of our knowledge, no previous study based on deep learning has been carried out to identify UC activity based on neutrophils detection using WSI, which has proven to be an accurate indicator of disease activity. In this work, we present a novel deep learning strategy to distinguish histological remission from activity based on the detection of neutrophils following the PHRI index. In summary, the main contributions of this work are:

- A deep learning framework used for the first time to accurately predict ulcerative colitis activity based on neutrophil detection.
- A novel constrained formulation that leverages prior knowledge in terms of relative tissue location (i.e. neutrophil location in the WSI) by imposing constraints on the feature extractor at bag (WSI)-level.
- A new attention weight for embedding-level MIL, which enlarges the relevance of the positive instances.
- We benchmark the proposed model against relevant body of literature on PICASSO-MIL, a large cohort of biopsies collected and digitalized in 7 centers in the UK, Germany, Belgium, Italy, Canada and USA.
- Comprehensive experiments demonstrate the superior performance of our model. By simply incorporating information about neutrophil location during the training, we found improvements of nearly 10% for bag-level classification compared to prior MIL methods.

2. Related work

2.1. Multiple instance learning

Multiple instance learning (MIL), a particular form of weakly-supervised learning, aims at training a model using a set of weakly labeled data [23]. In MIL tasks, the training dataset is composed of bags, where each one contains a set of instances and its goal is to teach a model to predict the bag label. A positive label is assigned to a bag if it contains at least one positive instance. MIL approaches have been successfully applied to computational histopathology for tasks such as tumor detection based on WSIs, reducing the time required to perform accurate annotations [24–29]. Some of these works use convolutional neural networks (CNNs) for the feature extraction process in each instance independently and then combine the instance-level information into one bag-level output. Methods that combine instance-level features are known as embedding-based, which require a later classification layer. In the case of [25], the bag level representation is achieved by the aggregation of the features through a simple batch global max-pooling (BGMP). Recent methods have proposed weighted-average embeddings, using instance-specific attention weights learned via a multi-layered perceptron projection or recurrent neural networks. In contrast, instance-based architectures combine instance-level predictions directly into the bag classification. In this vein, [24] obtained a tile-level feature representation through a CNN. These representations were then used in a recurrent neural network to integrate the information across the whole slide and report the final classification result to obtain a final slide-level diagnosis.

In most MIL-based papers, the WSIs employed have broad features that determine that a bag is positive. However, in this case,

* Corresponding author.
E-mail address: madeam2@upvnet.upv.es (R. del Amor).

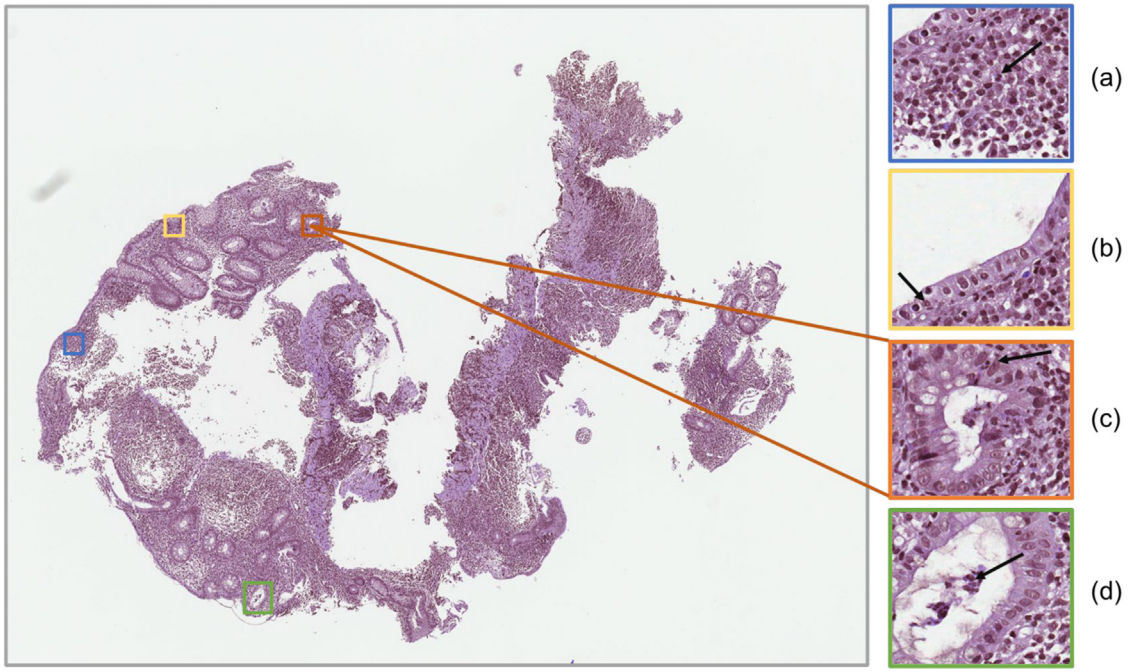

Fig. 1. The larger image corresponds to a Whole-Slide Image (WSI) of a patient suffering from ulcerative colitis. The patches marked with colours denote different interest structures. Specifically: (a) lamina propia, (b) surface epithelium, (c) cryptal epithelium and (d) cryptal lumen. The black mark indicates the presence of a neutrophil.

small cells (neutrophils) with features very similar to others in the tissue differentiate whether a bag is positive. Therefore, the typical MIL approach is not useful as the extracted activations are degrade and do not allow satisfactory classification.

2.2. Constrained CNNs

Constrained classification aims to guide the training of a CNN towards a solution that satisfies a given condition, which takes advantage of additional knowledge to the global labels. This learning paradigm has gained popularity on weakly-supervised scenarios (e.g. weakly supervised segmentation or MIL) since it allows to incorporate local information for improving the final task. Several works have tackled the problem of weakly-supervised segmentation by imposing constraints on deep CNNs [30–33]. In [30], the authors proposed a latent distribution and KL-divergence to constrain the output of a segmentation network. It is used in a semi-supervised setting to impose size constraints and image-level tags (i.e., force the presence or absence of given labels) on the regions of unlabeled images. Moreover, an L2 penalty term was proposed in [31] to impose equality constraints on the size of the target regions in the context of histopathology image segmentation which considerably improved the results. More recently, the authors showed in [32] that imposing inequality constraints on size directly in gradient-based optimization, also via an L2 penalty term, provided better accuracy and stability when few pixels of an image are labeled. Similarly, Zhou et al. embedded prior knowledge on the target size in the loss function by matching the probabilities of the empirical and predicted output distributions via the KL divergence. As directly minimizing this term by standard SGD is difficult, they proposed to optimize it by using stochastic primal-dual gradient [33]. While these works have helped to improve segmentation in a weakly-supervised setting, few studies focused on classification frameworks. In this work, by means of location constraints, we force the activations of the feature extractor to focus on those regions where neutrophils are localized. In this way, a reduced number of annotations can significantly improve the classification results.

3. Methodology

Here, we build an end-to-end MIL method as our baseline to perform image-to-image learning and prediction. The MIL formulation, based on CNNs, enables to detect neutrophils in WSIs and classify them into either histological remission or adverse outcome (UC activity). In Fig. 2, the proposed framework is shown. In the following, we describe the problem formulation and each of the proposed components.

3.1. Problem formulation

In MIL tasks, the training dataset is composed of bags, where each bag contains a set of instances (patches). A positive label is assigned to a bag if it has at least one positive instance. The goal of MIL is to teach a model to predict the bag label.

We denote our training dataset by $\mathcal{S} = (X_k, Y_k)$ with $k = \{1, 2, 3, \ldots, N\}$, where X_k denotes the kth input bag (WSI) and $Y_k \in 0, 1$ refers to the global label (ground truth label) assigned to the kth input WSI. Here, $Y_k = 0$ refers to a WSI with remission and $Y_k = 1$ refers to ulcerative colitis activity. Note that we denote each individual bag or WSI as: $X_k = \{x_{k,1}, \ldots, x_{k,t}, x_{k,I_n}\}$, where $x_{k,t}$ is the t-th instance of the bag and I_n denotes the total number of patches or instances in a slide. The number of instances varies considerably between slides.

The loss function used to optimize the end-to-end MIL approach is the cross-entropy cost function:

$$\mathcal{L}_{mil} = \sum_k (I(Y_k = 1) log\hat{Y}_k + I(Y_k = 0) log(1 - \hat{Y}_k) \quad (1)$$

where $I(.)$ is an indicator function.

3.2. MIL backbone with location constraints

As will be shown in the experiment section, our baseline MIL formulation produces a decent result for the proposed task but still with room for improvement. One problem is that the positive instances predicted by the algorithm tend to outgrow the true

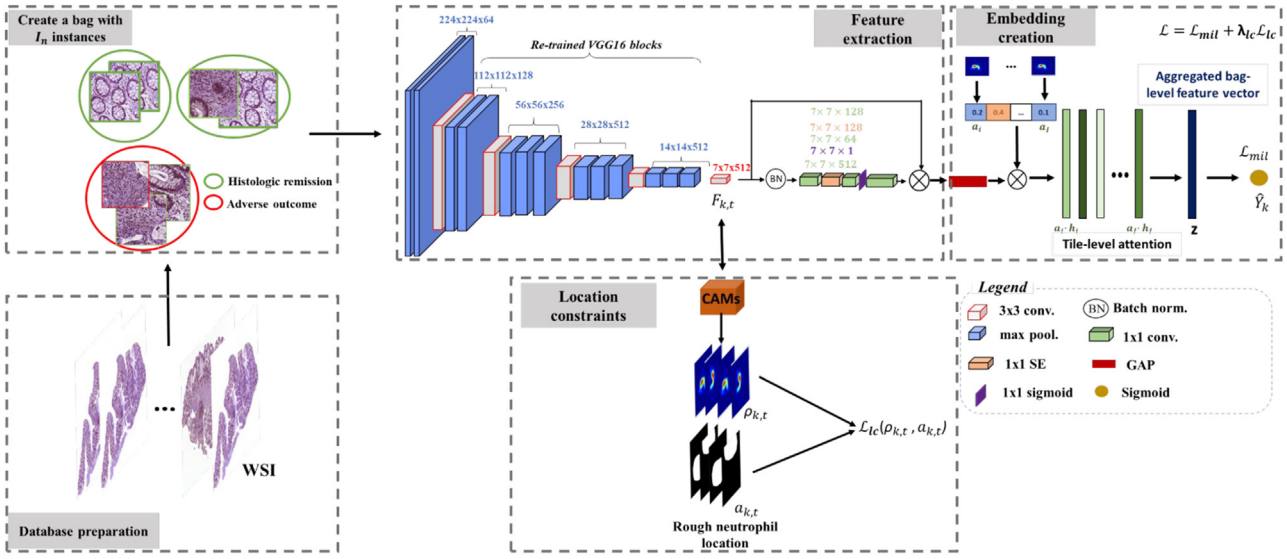

Fig. 2. Pipeline showing the embedded-level approach for ulcerative colitis detection. By incorporating the proposed location constraints, we force the backbone to extract more significant features from each patch belonging to a given bag. After that, we classify the entire biopsy using an aggregated bag-level feature vector weighted by the proposed attention-embedding weights.

regions with inflammation (UC activity) progressively. We propose using a neutrophil area constraint term to restrict the expansion of positive instances during training. We refer to our algorithm as location constrained MIL, abbreviated as LCMIL.

We denote our training set as $\mathcal{S} = (X_k, Y_k, A_k)$ with $k = \{1, 2, 3, \ldots, N\}$, where X_k denotes the kth bag, $Y_k \in \{0, 1\}$ refers to the global label (ground truth label) assigned to the kth input WSI and A_k specifies a rough estimation of the relative area in which the neutrophils are located within the image X_k. Being $a(i, j)_{k,t}$ the pixel (i, j) in the t-th patch from the bag k-th, $a(i, j)_{k,t} = 1$ if it corresponds to a pixel that is located around a neutrophil, whereas $a(i, j)_{k,t} = 0$, otherwise. Note that the rough annotations of neutrophil areas only are used for optimizing the parameters of the network (θ) and not for the prediction phase.

A Global-aggregation layer is implemented to obtain an activation map representing the distribution of the features extracted from each of the instances belonging to a given bag. This layer summarizes the information from all spatial locations in the feature-embedded map $F_{k,t} \in \mathbb{R}^{H \times W \times C}$ (corresponding to the last volume of features extracted by the backbone) to one representative map $\rho \in \mathbb{R}^{H \times W}$. Note that $H \times W$ are the dimensions of the instances and C is the number of filters. Therefore, $\rho \in \mathbb{R}^{H \times W}$ is defined as follows:

$$\rho(i, j)_{k,t} = \frac{1}{C} \sum_{c \in C} F_{k,t}(i, j, c) \tag{2}$$

In this way, we have a representation of how the backbone attention is distributed over the instance surface. In order to have the same dimension as the input instances (224^2), a bilinear interpolation is performed to the activation map ρ. In the following step, ρ is transformed into $\rho_s = \phi(\rho)$, where ϕ is the sigmoid activation function. The aim of the sigmoid activation function is to range the map activation function into [0–1]. Then, we define an area constraint as the L_2 penalty:

$$\mathcal{L}_{lc} = \sum_{k,t} I(Y_k = 1 \ and \ a(ij)_{k,t} > 0) \ ((a_{k,t} - \phi(\rho_{k,t}))^2) \tag{3}$$

Naturally, the global loss function can be updated from Eq. (1) to:

$$\mathcal{L} = \mathcal{L}_{mil} + \lambda_{lc}\mathcal{L}_{lc} \tag{4}$$

where $\lambda_{lc} \in \mathbb{R}^+$ weights the importance of the constraint during training.

3.3. MIL attention-embedding weights

After the feature extraction of each instance, we obtain a C-dimensional feature vector. The bag label predictor is in charge of aggregating the C-dimensional feature vectors $\{\mathbf{h}_t\}_{t \in I_n}$ into an embedding vector $Z_k \in \mathbb{R}^{1 \times C}$ representative of each bag. In the literature, there exist different simple aggregation functions such as batch global max-pooling (BGMP) or batch global average pooling (BGAP). However, these operators have a clear disadvantage. They are pre-defined and non-trainable. Other works use trainable aggregation functions [34]. However, in some situations, these attention weights have the same value for all instances in the bag, which is not suitable to determine a positive bag. This could be due to the complexity of the instance in some bags and the overfitting tendency of neural networks. To solve this problem, we propose to use a weighted average of instances where weights are obtained from the representative maps $\rho_{k,t}$. Note that the weights of these maps are updated each epoch using the $\mathcal{L}_{lc}$ term. Additionally, the weights must sum to 1 to be invariant to the size of a bag.

Therefore, the embedded feature vector per bag is obtained as $Z_k = \sum_{t \in I_n} a_t \cdot \mathbf{h}_t$, where a_t is defined as:

$$a_t = \frac{exp\{\sum \rho(i, j)/S\}}{\sum_{I_n} exp\{\sum \rho(i, j)/S\}} \tag{5}$$

where $S = H \cdot W$.

This attention vector promotes variability between instances of a positive bag. If there is no activation corresponding to neutrophils in the map ($\rho_{k,t}$), the value of a_t will be low and therefore, the embedding features $\mathbf{h}_t$ will have smaller weight in the final prediction. In the case of a negative bag, the attention values will be very similar and all instances will contribute equally. The superiority of this aggregation function for neutrophil identification and HR prediction will be shown in Section 4.

Table 2
Database description. Amount of whole-slide images (first row), number of patches (second row) and percentage of slides with PHRI>0, ulcerative colitis (third row).

	Training	Validation	Test
Number of WSI	84 (64,6%)	46 (35,4%)	100
patches	61.1 ± 54.2	58.2 ± 36.4	481.2 ± 292.1
PHRI score>0	51,1 %	39,15%	48%

4. Experiments and results

4.1. Implementation

All the tested approaches were implemented using Tensorflow 2.3.1 with Python. Experiments were conducted on the NVIDIA DGXA100 system.

1) **Dataset (PICASSO-MIL)**: We analyzed 230 colorectal biopsies from UC patients enrolled in a prospective international multicenter study to evaluate the proposed deep-learning methodology. Note that the slides belong to 7 different hospitals [35]. To process the large WSIs, these were downsampled to 20x resolution, divided into patches of size 512x512x3 with a 50% overlap among them. Aiming at pre-processing the biopsies and reducing the noisy patches, a mask indicating the presence of tissue in the patches was obtained by applying the Otsu threshold method over the magenta channel. Subsequently, the patches with less than 20% of tissue were excluded from the database. Using this database, we carried out a patient-level data partitioning procedure to separate training and validation sets, aiming to avoid overestimating the system's performance and ensuring its ability to generalize. Additionally, 100 non-annotated images at pixel-level were used to test the framework, see Table 2. During training, the human pathologists (with more than 35-year clinical experience) make two image-level annotations for each WSI, indicating each image as HR or UC activity depending on PHRI, and roughly estimating which areas of the image show neutrophils and inflammation. Only the bag label is necessary to evaluate the proposed method.

2) **Model parameters**: The MIL loss is known to be hard to train and special care is required for choosing training hyperparameters. To reduce fluctuations in optimizing the MIL loss, all training data are used in each iteration (the minibatch size is equal to the size of the training set). The network is trained with stochastic gradient descent (SGD) optimizer and a fixed learning rate of 0.01. The number of epochs was adapted in function of the experiment performed.

3) **Backbone network**: We choose the SeaNet (with VGG16) proposed in [36] as the CNN architecture of our framework since it demonstrated the improvement over standard methods in histological imaging. This framework is composed of VGG16 as a feature extractor and a squeeze and excitation attention network. In addition, we performed fine-tuning of this model, as it had previously been trained with histological images, in a different task, the detection of skin tumors.

4) **Evaluation**: The quantitative comparison of the different methodologies was handled by means of different figures of merit, such as sensitivity (SN), specificity (SPC), positive predictive value (PPV), false-positive rate (FPR) negative predictive value (NPV), F1-score (F1S), accuracy (ACC) and area under the ROC Curve (AUC).

4.2. Ablation experiments

In the following, we provide comprehensive ablation experiments to validate several elements of our model (LCMIL), and motivate the choice of the values employed in our formulation, as well as our experimental setting.

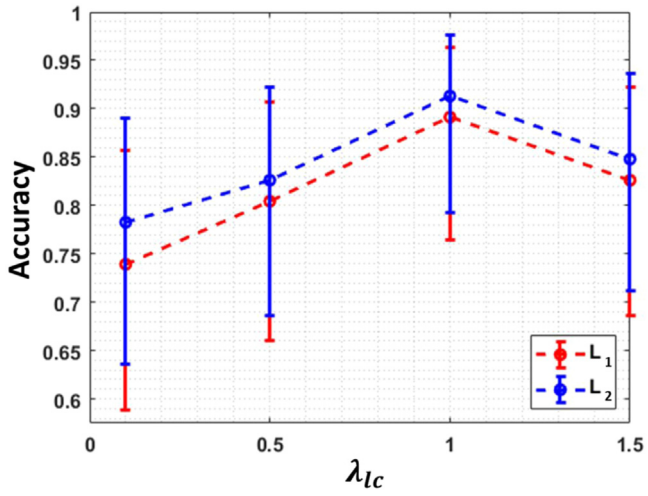

Fig. 3. Ablation studies on MIL formulation. Hyperparameters study for λ_{ac} are performed for bag-level accuracy on validation set. Confidence intervals are shown at 95%.

1) **Weight of location constraint loss**: The weight of the constraint loss is crucial for LCMIL since it directly decides the strength of constraints. Strong constraints may make the network unable to converge, while weak constraints have little help with learning. Therefore, we optimized the proposed formulation with the location constraint term in Eq. (4). Using the training setting previously described, we cross-validated different values of $\lambda_{ac} = \{0.1, 0.1, 1, 1, 5\}$. Additionally, we tried two loss functions, $\mathcal{L}_1$ and $\mathcal{L}_2$, to check for differences. We obtained bag-level ACC from the validation subset using the ACC on validation subset as early stopping criteria. Results are presented in Fig. 3.

These results show that the inclusion of the $\mathcal{L}_{lc}$ term improves the performance at bag level. Nevertheless, using a too large slope once the performance is satisfied can lead to a worsening of the results. Thus, we selected $\lambda_{lc} = 1$, which led to the best results at bag level in the validation cohort.

Additionally, we want to get a more intuitive view of how the proposed methodology location constraint term influences the extraction of discriminative features. For that purpose, we depict the feature representation of the embedding space produced by the encoder networks of MIL without $\mathcal{L}_{lc}$ and the proposed encoder on the instance-level labeled validation. Concretely, we obtained the class activation maps for regions of a bag where neutrophils are found (cryptal lumen, cryptal epithelium, lamina propia and surface epithelium). In Fig. 4, the annotations made by the pathologists, the activation maps obtained by a MIL module without $\mathcal{L}_{lc}$ and the proposed method are compared.

The MIL without location constraint module does not focus its attention on the areas where neutrophils are located by the pathologist but on other cells found in the tissue. Note that neutrophils are very similar to other cells found in the tissue, such as eosinophils, macrophages, etc., but in this case, they do not determine that a patient has active ulcerative colitis. This is why the specificity of this model is very low. In contrast, the inclusion of the location constraints module forces the network to focus its attention on the real determining cells, the neutrophils. In this way, we can therefore obtain precise instance-level maps for unannotated images that allow us to detect the neutrophils.

2) **Attention weights for bag classification**: Using the best configuration reached for the λ_{lc} term, we optimized the embedded feature vector per bag, see Table 3. This Table compares the best-known methodologies for constructing the embedded vector (BGAP, BGMP and MIL-Attention) versus the proposed method.

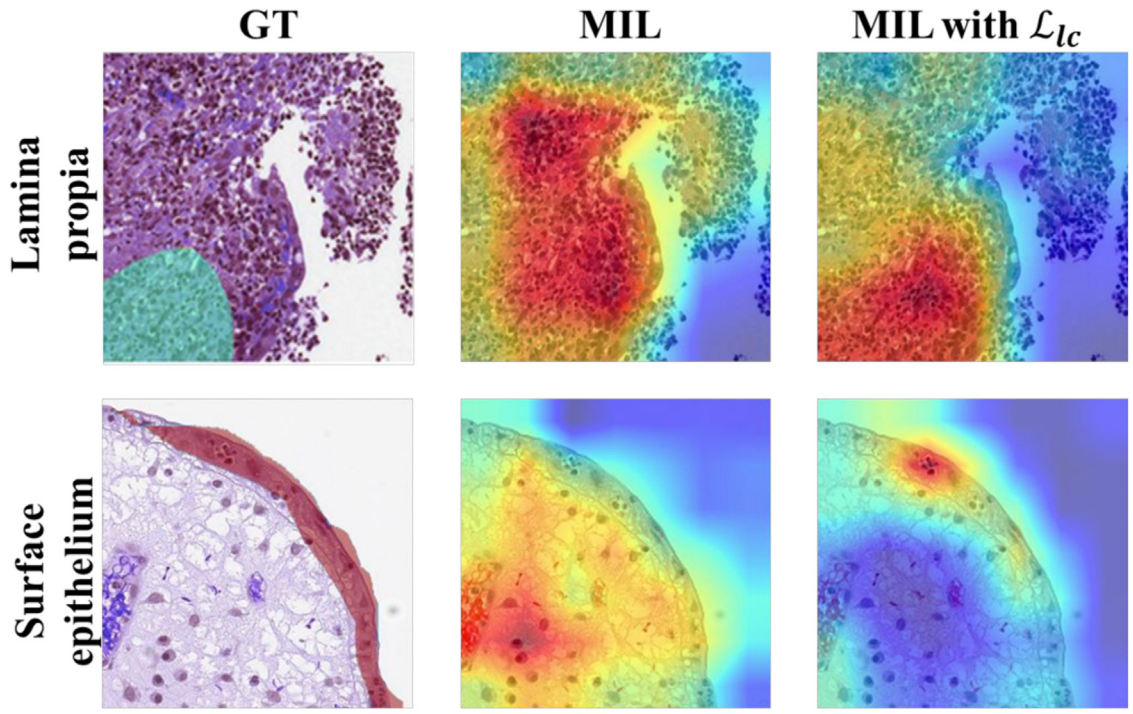

Fig. 4. Class activation maps (CAMs) of some regions where neuthophils are found. First column: original images with pathologist annotation (green and red annotations); Second column: CAMs obtained using the normal MIL model. Third column: CAMs using the proposed location constraints.

Table 3
Comparison of the different attention embedding weights on the validation set. BGAP: batch global average pooling, BGMP: batch global max-pooling, LCMIL: neutrophil constrained weak supervision (proposed). Note that in all cases the location constraint proposed is integrated into the backbone.

	BGAP	BGMP	**Attention** [34]	LCMIL
SN	0.9643	0.9643	0.8889	**0.9643**
SPC	0.6667	0.7778	0.7778	**0.8333**
PPV	0.8182	0.8710	0.8571	**0.9000**
NPV	0.9231	0.9333	0.8235	**0.9375**
F1S	0.8852	0.9153	0.8727	**0.9310**
ACC	0.8478	0.8913	0.8444	**0.9130**
AUC	0.8155	0.8710	0.8333	**0.8988**

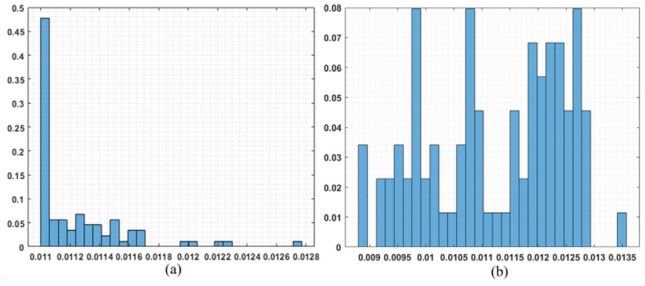

Fig. 5. Distribution of embedding weights across the instances that comprise a WSI. (a) Proposed attention embeddings. (b) Attention weights proposed in [34].

Since the features that discriminate a positive bag are relatively small compared to the dimension of the different instances, in this case, the BGMP layer improves the results of the BGAP and MIL-Attention layers. However, the proposed aggregation method outperforms all previous methods.

To compare the distribution of the attention weights of [34] with those proposed here, we show the histogram of these values in a positive bag, see Fig. 5. In this case, the bag comprises 80 instances, of which only 15% are positive, i.e., contain neutrophil structures. In Fig. 5(b), attention proposed in [34], the different values of weights have similar probabilities. Therefore, no discriminatory weighting is performed to separate negative and positive instances. However, with the proposed method, most instances (around 60) have a low weight, which would belong to the instances without neutrophils. The remaining weights are spread across instances with neutrophils, with higher weights assigned to those with more significant features. Therefore, the proposed attention-based MIL allows to assign more discriminate weights to instances within a bag and hence the final representation of the bag is highly informative for the bag-level classifier.

Table 4
Comparison of the different baseline frameworks in the test cohort. Note that for the test cohort only the global bag label are available.

	ABMIL	DSMIL	CLAM-SB	MIL-RNN	LCMIL
SN	0.9583	0.8293	0.9302	0.8667	**0.9583**
SPC	0.6923	0.7288	0.8033	0.7797	**0.9615**
PPV	0.7419	0.6800	0.7692	0.7500	**0.9583**
NPV	0.9473	0.8600	0.9423	0.8846	**0.9615**
F1S	0.8393	0.7473	0.8421	0.8041	**0.9583**
ACC	0.8200	0.7700	0.8558	0.8173	**0.9600**
AUC	0.8253	0.7546	0.8321	0.8009	**0.9599**

4.3. Comparison to the literature

To compare the proposed method with the MIL baselines, a comparative analysis of the test cohort is performed in this section, see Table 4. For this purpose, we included the current state-of-the-art deep MIL models, the attention based pooling operator (ABMIL) [34], non-local attention based pooling operator (DSMIL) [28], single-attention-branch (CLAM-SB) [29] and recurrent neural network (RNN) based aggregation (MIL-RNN) [24].

The figures of merit are obtained at the biopsy label because only these labels are available in the test set. In general, the specificity of the MIL baseline models drops considerably. The best

state-of-the-art model (CLAM-SB) achieves a specificity of 0.8033 compared to 0.9615 obtained by the proposed model (LCMIL). State-of-the-art models are not able to discriminate between neutrophils and other tissue cells and therefore are not optimal for predicting diseases such as ulcerative colitis, which are caused by very precise histological patterns. Under our proposed formulation (LCMIL), the model can detect neutrophils at the instance level and, therefore, predicts ulcerative colitis with a good performance. Obviously, there is a high consistency between the fine annotation area and CAMs obtained in Fig. 4, illustrating great interpretability and attention visualization of the proposed framework. Therefore, with a small volume of training annotations, the model can improve the accuracy of the best baseline MIL approach by almost 10%.

5. Conclusion

Whole-slide images (WSI) have shown applicability to developing computer vision models, but few studies have approached the use of deep learning models to detect ulcerative colitis (UC). In this work, we propose an location constraint framework able to perform histological remission prediction using WSIs of patients with UC. Our framework comprises a feature extraction backbone with an attention module to refine the patch-level features and a MIL approach to predict the UC activity in each bag. We introduce a location constraint module that forces the feature extractor to focus on the most significant patterns in the patches that form a bag. The biopsy classification comes from the bag-level feature vector that the attention embedding has ponderated. This approach reaches a test accuracy of 0.9600 in a more significant subset than the training set, which shows that the extra pixel-level annotation gives crucial information to the algorithm.

Future research lines need to focus on detecting neutrophils in the different biopsy regions and grading PHRI accordingly, not being limited to the histological activity or remission prediction. The location constraint approach also promises applicability to other pathologists in which histological analysis is based on identifying single cells.

Funding

This work has received funding from Horizon 2020, the European Union's Framework Programme for Research and Innovation, under grant agreement No. 860627 (CLARIFY), the Spanish Ministry of Economy and Competitiveness through project PID2019-105142RB-C21 (AI4SKIN) and GVA through projects PROMETEO/2019/109 and INNEST/2021/321 (SAMUEL). Rocío del Amor and Adrián Colomer work have also been supported by the Spanish Government under FPU Grant (FPU20/05263) and the Universitat Politécnica de Valéncia (PAID-10-21 - Subprograma 1), respectively.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgements

We gratefully acknowledge the support from the Generalitat Valenciana (GVA) with the donation of the DGX A100 used for this work, action co-financed by the European Union through the Operational Program of the European Regional Development Fund of the Comunitat Valenciana 2014–2020 (IDIFEDER/2020/030).

References

[1] S.C. Ng, H.Y. Shi, N. Hamidi, F.E. Underwood, W. Tang, E.I. Benchimol, R. Panaccione, S. Ghosh, J.C. Wu, F.K. Chan, et al., Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies, Lancet 390 (10114) (2017) 2769–2778.
[2] D. Jain, B.F. Warren, R.H. Riddell, Inflammatory disorders of the large intestine, in: Morson and Dawson's Gastrointestinal Pathology, 2013, pp. 552–635.
[3] R.V. Bryant, D.C. Burger, J. Delo, A.J. Walsh, S. Thomas, A. von Herbay, O.C. Buchel, L. White, O. Brain, S. Keshav, et al., Beyond endoscopic mucosal healing in UC: histological remission better predicts corticosteroid use and hospitalisation over 6 years of follow-up, Gut 65 (3) (2016) 408–414.
[4] V. Narang, R. Kaur, B. Garg, R. Mahajan, V. Midha, N. Sood, A. Sood, Association of endoscopic and histological remission with clinical course in patients of ulcerative colitis, Intest Res. 16 (1) (2018) 55.
[5] A. Ponte, R. Pinho, S. Fernandes, A. Rodrigues, L. Alberto, J.C. Silva, J. Silva, J. Rodrigues, M. Sousa, A.P. Silva, et al., Impact of histological and endoscopic remissions on clinical recurrence and recurrence-free time in ulcerative colitis, Inflamm. Bowel Dis. 23 (12) (2017) 2238–2244.
[6] T. Lobatón, T. Bessissow, A. Ruiz-Cerulla, G. De Hertogh, R. Bisschops, J. Guardiola, G. Van Assche, S. Vermeire, M. Ferrante, Prognostic value of histological activity in patients with ulcerative colitis in deep remission: a prospective multicenter study, United European Gastroenterol. J. 6 (5) (2018) 765–772.
[7] T.E. Römkens, P. Kranenburg, A.v. Tilburg, C. Bronkhorst, I.D. Nagtegaal, J.P. Drenth, F. Hoentjen, Assessment of histological remission in ulcerative colitis: discrepancies between daily practice and expert opinion, J. Crohn's Colitis 12 (4) (2018) 425–431.
[8] D. Alsoud, G. Compernolle, S. Tops, J. Sabino, M. Ferrante, D. Thomas, G. De Hertogh, S. Vermeire, B. Verstockt, P442 real-world endoscopic and histologic outcomes are linked to ustekinumab exposure in ulcerative colitis, J. Crohn's Colitis 16 (Supplement_1) (2022). i424–i424
[9] F. Magro, J. Lopes, P. Borralho, S. Lopes, R. Coelho, J. Cotter, F.D. de Castro, H.T. de Sousa, M. Salgado, P. Andrade, et al., Comparison of different histological indexes in the assessment of UC activity and their accuracy regarding endoscopic outcomes and faecal calprotectin levels, Gut 68 (4) (2019) 594–603.
[10] M.H. Mosli, B.G. Feagan, W.J. Sandborn, G. D'Haens, C. Behling, K. Kaplan, D.K. Driman, L.M. Shackelton, K.A. Baker, J.K. MacDonald, et al., Histologic evaluation of ulcerative colitis: a systematic review of disease activity indices, Inflamm. Bowel Dis. 20 (3) (2014) 564–575.
[11] A. Mojtahed, R. Khanna, W.J. Sandborn, G.R. D'Haens, B.G. Feagan, L.M. Shackelton, K.A. Baker, E. Dubcenco, M.A. Valasek, K. Geboes, et al., Assessment of histologic disease activity in Crohn's disease: a systematic review, Inflamm. Bowel Dis. 20 (11) (2014) 2092–2103.
[12] S. Riley, V. Mani, M. Goodman, S. Dutt, M. Herd, Microscopic activity in ulcerative colitis: what does it mean? Gut 32 (2) (1991) 174–178.
[13] R.K. Pai, D.J. Hartman, C.R. Rivers, M. Regueiro, M. Schwartz, D.G. Binion, R.K. Pai, Complete resolution of mucosal neutrophils associates with improved long-term clinical outcomes of patients with ulcerative colitis, Clin. Gastroenterol. Hepatol. 18 (11) (2020) 2510–2517.
[14] F. Magro, G. Doherty, L. Peyrin-Biroulet, M. Svrcek, P. Borralho, A. Walsh, F. Carneiro, F. Rosini, G. de Hertogh, L. Biedermann, et al., ECCO position paper: harmonization of the approach to ulcerative colitis histopathology, J. Crohn's Colitis 14 (11) (2020) 1503–1511.
[15] C. Ma, R. Sedano, A. Almradi, N.V. Casteele, C.E. Parker, L. Guizzetti, D.F. Schaeffer, R.H. Riddell, R.K. Pai, R. Battat, et al., An international consensus to standardize integration of histopathology in ulcerative colitis clinical trials, Gastroenterology 160 (7) (2021) 2291–2302.
[16] X. Gui, A. Pathology Bazarova, R. del Amor, M. Vieth, G. de Hertogh, V. Villanacci, D. Zardo, T. Parigi, E. Røyset, U. Shivaji, et al., PICaSSO histologic remission index (PHRI) in ulcerative colitis–development of a novel simplified histological score for monitoring mucosal healing and predicting clinical outcomes and its applicability in an artificial intelligence system, Gut (2022).
[17] R.W. Stidham, W. Liu, S. Bishu, M.D. Rice, P.D. Higgins, J. Zhu, B.K. Nallamothu, A.K. Waljee, Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis, JAMA Netw. Open 2 (5) (2019). e193963–e193963
[18] T. Ozawa, S. Ishihara, M. Fujishiro, H. Saito, Y. Kumagai, S. Shichijo, K. Aoyama, T. Tada, Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis, Gastrointest. Endosc. 89 (2) (2019) 416–421.
[19] A. Alammari, A.R. Islam, J. Oh, W. Tavanapong, J. Wong, P.C. De Groen, Classification of ulcerative colitis severity in colonoscopy videos using CNN, in: Proceedings of the 9th International Conference on Information Management and Engineering, 2017, pp. 139–144.
[20] M. Byrne, J. East, M. Iacucci, R. Panaccione, R. Kalapala, N. Duvvur, H. Rughwani, A. Singh, M. Henkel, S. Berry, et al., DOP13 artificial intelligence (AI) in endoscopy-deep learning for detection and scoring of ulcerative colitis (UC) disease activity under multiple scoring systems, J. Crohn's Colitis 15 (Supplement_1) (2021) S051–S052.
[21] H.P. Bhambhvani, A. Zamora, Deep learning enabled classification of mayo endoscopic subscore in patients with ulcerative colitis, Eur. J. Gastroenterol. Hepatol. 33 (5) (2021) 645–649.
[22] V. Casteele, J.A. Leighton, S.F. Pasha, F. Cusimano, A. Mookhoek, C.E. Hagen, C. Rosty, R.K. Pai, Utilizing deep learning to analyze whole slide images of colonic biopsies for associations between eosinophil density and clinicopathologic features in active ulcerative colitis, Inflamm. Bowel Dis. (2021).
[23] C.L. Srinidhi, O. Ciga, A.L. Martel, Deep neural network models for computational histopathology: a survey, Med. Image Anal. (2020) 101813.

[24] G. Campanella, M.G. Hanna, L. Geneslaw, A. Miraflor, V.W.K. Silva, K.J. Busam, E. Brogi, V.E. Reuter, D.S. Klimstra, T.J. Fuchs, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nat. Med. 25 (8) (2019) 1301–1309.
[25] K. Das, S. Conjeti, J. Chatterjee, D. Sheet, Detection of breast cancer from whole slide histopathological images using deep multiple instance CNN, IEEE Access (2020) 213502–213511.
[26] Y. Zhao, F. Yang, Y. Fang, H. Liu, N. Zhou, J. Zhang, J. Sun, S. Yang, B. Menze, X. Fan, et al., Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4837–4846.
[27] J. Silva-Rodriguez, A. Colomer, J. Dolz, V. Naranjo, Self-learning for weakly supervised Gleason grading of local patterns, IEEE J. Biomed. Health Inform. (2021) 3094–3104.
[28] B. Li, Y. Li, K.W. Eliceiri, Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14318–14328.
[29] M.Y. Lu, D.F. Williamson, T.Y. Chen, R.J. Chen, M. Barbieri, F. Mahmood, Data-efficient and weakly supervised computational pathology on whole-slide images, Nat. Biomed. Eng. 5 (6) (2021) 555–570.
[30] D. Pathak, P. Krahenbuhl, T. Darrell, Constrained convolutional neural networks for weakly supervised segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1796–1804.
[31] Z. Jia, X. Huang, I. Eric, C. Chang, Y. Xu, Constrained deep weak supervision for histopathology image segmentation, IEEE Trans. Med. Imaging 36 (11) (2017) 2376–2388.
[32] H. Kervadec, J. Dolz, M. Tang, E. Granger, Y. Boykov, I.B. Ayed, Constrained-CNN losses for weakly supervised segmentation, Med. Image Anal. 54 (2019) 88–99.
[33] Y. Zhu, S. Tang, L. Quan, W. Jiang, L. Zhou, Extraction method for signal effective component based on extreme-point symmetric mode decomposition and Kullback–Leibler divergence, J. Br. Soc. Mech. Sci. Eng. 41 (2) (2019) 1–11.
[34] M. Ilse, J. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: International Conference on Machine Learning, 2018, pp. 2127–2136.
[35] M. Iacucci, S.C. Smith, A. Bazarova, U.N. Shivaji, P. Bhandari, R. Cannatelli, M. Daperno, J. Ferraz, M. Goetz, X. Gui, et al., An international multicenter real-life prospective study of electronic chromoendoscopy score picasso in ulcerative colitis, Gastroenterology 160 (5) (2021) 1558–1569.
[36] R. Del Amor, L. Launet, A. Colomer, A. Moscardó, A. Mosquera-Zamudio, C. Monteagudo, V. Naranjo, An attention-based weakly supervised framework for spitzoid melanocytic lesion diagnosis in whole slide images, Artif. Intell. Med. 121 (2021) 102197.