Classification of Volumetric Images Using Multi-Instance Learning and Extreme Value Theorem

Ruwan Tennakoon, Gerda Bortsova, Silas Ørting, Amirali K. Gostar, Mathilde M. W. Wille, Zaigham Saghir, Reza Hoseinnezhad, Marleen de Bruijne, and Alireza Bab-Hadiashar.

***Abstract*—Volumetric imaging is an essential diagnostic tool for medical practitioners. The use of popular techniques such as convolutional neural networks (CNN) for analysis of volumetric images is constrained by the availability of detailed (with local annotations) training data and GPU memory. In this paper, the volumetric image classification problem is posed as a multi-instance classification problem and a novel method is proposed to adaptively select positive instances from positive bags during the training phase. This method uses the extreme value theory to model the feature distribution of the images without a pathology and use it to identify positive instances of an imaged pathology. The experimental results, on three separate image classification tasks (i.e. classify retinal OCT images according to the presence or absence of fluid build-ups, emphysema detection in pulmonary 3D-CT images and detection of cancerous regions in 2D histopathology images) show that the proposed method produces classifiers that have similar performance to fully supervised methods and achieves the state of the art performance in all examined test cases.**

***Index Terms*—Multiple instance learning,weakly supervised learning, CNN, OCT, CT, Macular Edema, Emphysema, COPD.**

I. Introduction

VOLUMETRIC medical imaging is frequently used for disease diagnostics [1], [2]. Due to high memory usage and limited availability of training data, applying state-of-the-art classification techniques such as deep convolutional neural network (CNN) to volumetric image classification is challenging [3]. A typical 3D medical image contains in the order of $512 \times 512 \times 400$ voxels and feeding such images directly to a CNN would result in large memory usage. One option, as used in natural image classification, is to down-sample the image. The fine details that are important for diagnosis are often contained in a small area and down-sampling can easily obscure those instances.

The second common approach to address the above issue is to partition the image into compact volumes (i.e. instances), and train a CNN using human expert annotation of each instance (analogous to object detection framework) [4], [5]. At inference, the outputs of the network can then be aggregated according to some rule or using another classifier to generate image level predictions. The winning team of the "Data Science Bowl 2017" [6] followed a variant of the above approach. However, in practice, getting localized expert annotations is an expensive task and relying on those annotations reduce the applicability of these techniques.

The third common approach is to consider the problem as a multi-instance classification problem. In this technique, an image is converted into a bag of instances (in our case, image patches) and each bag has a label. A model that classify the input based on preesnce or absence of specific abnormalities can then be learned using the multiple instance learning (MIL) assumption [7]: At least one instance in a positive bag is positive and none of the instances in negative bags are positive. A recent survey of the use of multiple instance learning in medical imaging is presented in [8]. It is said that when the above formulation is used for learning classifiers, utilized information from the positive images is likely to be limited to a single instance [9]. In medical image applications with limited training data, such behaviour would not be desirable.

There have been several works that aim to include more than one instance from a positive bag during training. The methods that are closely related to the proposed method are discussed below. In simple-MIL approach [10], [11], [12] bag labels are propagated to instance labels and all the instances are used in training an instance classifier. Similarly, in mi-SVM [13], instance labels are initialized with bag labels and then updated iteratively in the training process by thresholding the classifier output. State-of-the-art examples of formulations that use more than one instance from a positive bag in learning CNN features include: 1) MIL based whole mammogram classification [9]: This work proposed to use the top-k instances from a positive bag as positive and the rest as negative. As the authors of the paper pointed out, defining a k that is appropriate for all images is challenging. As a possible solution, the authors proposed a soft method (i.e. sparse multi-instance learning),

R. Tennakoon, A. K. Gostar, R Hoseinnezhad and A. Bab-Hadiashar are with School of Engineering, RMIT University, Melbourne, Australia.

G Bortsova is with Biomedical Imaging Group Rotterdam, Department of Radiology and Medical Informatics, Erasmus MC, Rotterdam, The Netherlands.

S. Ørting is with Department of Computer Science, University of Copenhagen, Copenhagen, Denmark.

M. M. W. Wille is with Department of Diagnostic Imaging, Bispebjerg Hospital, Copenhagen, Denmark.

Z. Saghir is with Department of Respiratory Medicine, Gentofte University Hospital, Hellerup, Denmark.

M. de Bruijne is with Department of Computer Science, University of Copenhagen, Copenhagen, Denmark and Biomedical Imaging Group Rotterdam, Department of Radiology and Medical Informatics, Erasmus MC, Rotterdam, The Netherlands.

which adds a sparsity constraint on the number of positive instances to the cost function. However, implementing such constraints is not feasible in problems where the full bag cannot be loaded to memory at once for back-propagation. 2) CNN based whole slide tissue image classification [14]: The method defines hidden variables to flag instances containing discriminative information and then uses an EM algorithm to infer both the hidden variable and the CNN parameters, iteratively. Similarly, inferring the hidden variable requires a threshold and the authors note that using a simple thresholding scheme would ignore the useful instances that fall near the decision boundary. As such, they proposed an elaborate thresholding scheme consisting of two neural networks, spatial smoothing and image (and class) level thresholds (again requiring tuned parameters) to address this.

For many medical imaging applications including detection of emphysema in CT and retinal fluid presence in Optical Coherence Tomography (OCT), defining thresholds indicating how many positive instances are likely to be in a given image, as required by some of the above-mentioned methods, is challenging. One such example is elaborated in Fig. 1 which shows the number of positive instances per image for the three tasks in OCT image classification dataset [15]. The number of positive instances per image in this dataset has large variation and defining a fixed threshold would not be useful. An instance in the OCT dataset is an image patch with dimensions $[512 \times 320 \times 1]$ and is labelled positive if there is at least one positive pixel (identified using ground truth annotations) within the patch. Therefore, the number of positive instances would vary with the severity of the disease.

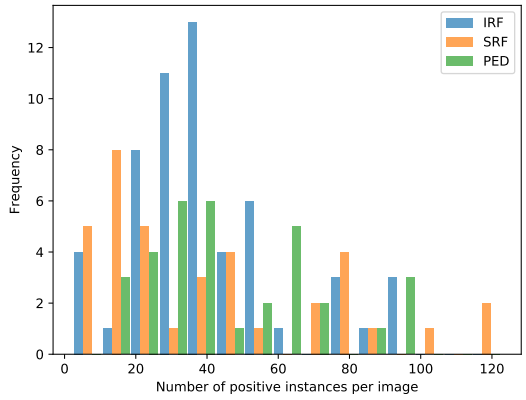


Fig. 1: Histograms of positive instances per image for three OCT image classification tasks [15]. The number of positive instances per image in this dataset has a large variation and cannot be properly segmented by a simple threshold.

In this paper, we propose to use extreme value theory (EVT) to model the maximum feature deviations (from the mean feature of negatives) of instances in the negative images and use this information to identify the probable positive instances in positive images. The proposed method eliminates the need for predefined thresholds and provides a mechanism for memory efficient end-to-end training of CNN in a weakly supervised setting. While extreme value theory has been used in computer vision in the past for SVM score calibration [16] and open-set classification [17], [18], its use for training weakly supervised multi-instance learning classifiers has not yet been explored.

The idea of using EVT to identify positive instances in a MIL setting was outlined in [19]. The current paper builds upon our earlier work by:

I Modifying the method to reduce the number of hyper-parameters and use a simplified cost function.
II Providing extended evaluations on multiple datasets from two different domains.
III Including detailed discussions highlighting the inner-workings of the proposed method.

The rest of this paper is organized as follows: Section II provides a brief overview of EVT and volumetric image classification using learning methods. Section III describes the proposed method while Section IV presents the experimental results on 3D OCT image classification and emphysema detection in 3D CT images. The inner-workings of the proposed method is examined in detail in Section V. Section VI concludes the paper.

II. Background

A. Deep-learning based volumetric image classification

Methods for deep-learning based volumetric image classification can be divided into three categories: 1) Patch based detection, 2) Unsupervised deep representation learning combined with traditional classifiers, 3) Multiple instance learning based methods.

In a typical patch based detection method, the training data consist of annotations of the pathology location and region. This information can be used to train CNN based classifiers that operate on smaller patches, thereby eliminating the high memory requirement at training time. At test time, all the patches in a test image are analyzed for pathology. Such framework has been employed in detecting pulmonary nodules [20], colonic polyps [21], and cerebral micro-bleeds [22].

Another possible framework to avoid whole image based training is to use unsupervised learning techniques to train a deep model (such as an auto-encoder) that can convert an input image patch to a low dimensional feature representation. The features extracted with this model for all the patches in a given image can be aggregated and used to learn a classifier with the corresponding whole image label [23]. Unsupervised learning might learn features that may not be relevant for the task at hand.

Multiple instance based deep learning methods [8] is a compromise between the above two approaches (that only use image level annotations at training time). These methods have been used in several applications including body part recognition [24] and mammogram classification [9] to name a few.

B. Multiple instance learning based classification

Multiple Instance Learning (MIL) is a variation of supervised learning where a label is only assigned to a collection

of observations or bag of instances. This poses additional challenges compared to standard supervised learning where each observation accompanies a label. Because the level of annotations required in MIL is significantly lower than supervised learning, it has attracted lot of attention in recent years particularly in areas including drug discovery, computer vision, text classification and signal processing [25]. Amores [25] categorized MIL based classification into three paradigms: Instance-space (IS), Bag-space (BS) and Embedded-space (ES).

In instance-space paradigm the discriminative information is assumed to be contained locally at instance level. The problem of classifying histopathology images containing cancerous regions is an example case where such assumptions is applicable [26]. As discriminative information exists locally at an instance level, an instance classifier can be trained using the standard MIL assumption and the output of this classifier for all instances in a bag can be aggregated to produce the bag label [13], [12], [27].

In both Bag and Embedded space paradigms, all instances in a bag has to be considered simultaneously because the discriminative information is assumed to lie globally. Triaging in diabetic retinopathy screening is an example of such problem [28], [29]. Bag-space methods aim to classify bags directly often by defining kernels or dissimilarity between bags [30], [31]. On the other hand, in ES paradigm, features corresponding to each instance in a bag is combined to produce a single embedding that is then used in a classifier to produce the bag label [27].

While traditional MIL methods use hand-crafted features to represent the instances and learn only the classifier, several methods have emerged in recent years that learn both the feature representation and the classifier simultaneously [27], [32]. The method proposed in this paper is an example of the instance-space paradigm and uses an end-to-end learning based framework. A detailed review of different MIL methods is provided in [33], [25], [8], [34].

C. Extreme value Distribution

Extreme value theorem (EVT) [35], which is commonly used for modelling unusual events in weather and financial systems, is a counterpart of the central limit theorem. While the central limit theorem describes the distribution of mean values, EVT describes the behaviour of extreme values sampled from any underlying distribution. To explain our proposed method, we first briefly explain the extreme value theorem [35]:

Theorem 1: For $M_n = \max(X_1, \ldots, X_n)$ where $(X_1, X_2, \ldots)$ are a sequence of i.i.d samples drawn from any distribution. If there exist a sequence of pairs of real numbers (a_n, b_n) with $a_n > 0 \ : \ \forall i$ such that

$$\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = G(x) \tag{1}$$

where G is a non-degenerate function, then G must be a member of one of the following distribution families: *Gumbel*, *Fréchet* and *Weibull*. ■

If samples are bounded from either side, the appropriate distribution of those extreme values will then be a *Weibull* distribution. The i.i.d assumption in the above theorem was later relaxed to a weaker assumption of exchangeable random variables in [36].

A more practical form of the above theorem is called the block maxima method of EVT and is used when there may exist local dependencies within blocks but not between blocks [37]. Our proposed modelling of negative instances falls into this category, where there might be some dependencies between the instances from a given image but it is reasonable to assume independence between images when they are of different subjects.

III. Proposed Method

Given a training dataset $\mathcal{X} = \left\{\left(X^{(i)}, y^{(i)}\right)\right\}_{i=1}^{N}$ containing N volumetric images, $X^{(i)} \in \mathbb{R}^{h_i \times w_i \times z_i}$, and the corresponding expert annotations, $y^{(i)} \in \{0, 1\}$ (indicating whether a particular abnormality is present or not), our intention is to learn a model that can predict whether the particular abnormality is present in an unseen image $X^{(\cdot)}$.

Our proposed method (EVT-MIL) is to use an iterative sampling based multiple-instance classification framework. In this approach each image is partitioned into a collection (bag) of smaller volumes (instances) $B^{(i)} := \{x_1^{(i)}, \ldots, x_{L_i}^{(i)}\}$; $x_l^{(i)} \subset X^{(i)}$ and, the probability that an abnormality is present in a particular instance is modelled using a parameterized function, $f_\theta\left(x_j^{(i)}\right) \Leftrightarrow P\left(h_j^{(i)} = 1 \mid x_j^{(i)}\right)$. Here $h_j^{(i)}$ is the instance label, that indicates the presence of abnormality in instance j of image i. Using the standard MIL assumption, the abnormality presence probability for the overall image can be inferred as follows:

$$p\left(y^{(\cdot)} = 1 \mid X^{(\cdot)}\right) = \max\left(f_\theta\left(x_1^{(\cdot)}\right), \ldots, f_\theta\left(x_{L_i}^{(\cdot)}\right)\right). \tag{2}$$

Now, the challenge is to learn the model parameters θ, given only the bag labels. We consider the instance labels as hidden variables and use an expectation-minimization (EM) based framework, shown in Fig. 2, to learn the model parameters θ. The remainder of this section describes the two main steps in

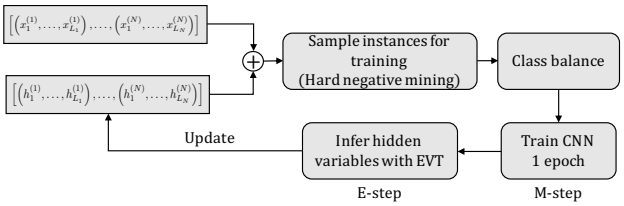


Fig. 2: The overall block diagram of the training phase.

the proposed EM based method: 1) Learning the model that maps an image instance to the current estimate of instance label h and 2) Inferring the values of h using the EVT.

A. Learning deep CNN model (M step)

In this work, the function f_θ is modelled using a CNN. The architecture of the CNN and the cost function used in the optimization step are as follows:

1) CNN architecture: The first stage of the network (i.e. base model) take raw voxels as input and produce a higher dimensional feature representation. In our experiments, we used the convolutional stages of the AlexNet architecture [38] for OCT segmentation and a 3D variant of Squeezenet architecture [39] for the emphysema detection, as the base model. The choice of base model is not restricted to the above architectures and any suitable CNN architecture can be used in this step. The details of used networks are provided with the code[1].

The output of the base model is connected to a fully connected structure through a global average pooling layer as shown in Fig. 3.

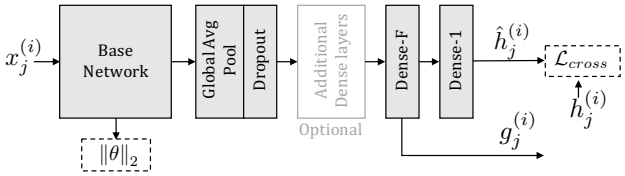


Fig. 3: Architecture of the proposed neural network model. Dense-F is the output of a fully connected layer with F units. The layers marked optional can be included depending on the task.

2) Cost function: The parameters of the overall network are obtained by minimizing a cost function consisting of two terms:

$$\theta^* = \arg\min_{\theta} \left\{ \mathcal{L}_{cr}(\hat{h}_j^{(i)}, h_j^{(i)}) + \lambda_1 \left\|\theta_b\right\|_2 \right\} \tag{3}$$

where λ_1 is a hyper-parameter that balance the terms in the cost function, $\|\theta_b\|_2$ is the L_2-norm of the base model parameters (acts as a regularizer) and $\mathcal{L}_{cr}$ is binary cross-entropy function. Weight regularization was not implemented in the last two fully connected layers, i.e. Dense-F, Dense-1.

B. Inferring instance labels with EVT (E step)

One of the main contributions of this paper is the novel way that labels are assigned to instances (inferring the hidden variables). Instead of thresholding the CNN output $\hat{h}_j^{(i)}$, as prescribed in [14], we propose to use the extreme value theorem to model extreme instances in the negative bags and then use that model to define the probability of being positive (i.e. positiveness) for instances in the positive bags. Under the MIL setting, it is assumed that all the instances in a negative bag are negative. Hence, the instances from negative bags can be used to approximate the distribution of negative instances. In our approach, use of a probabilistic model (instead of a deterministic one such as a fixed threshold) makes the classifier robust to influences of some incorrectly labelled bags. Furthermore, Using EVT provides a much sharper cumulative distribution function (CDF) and does not rely on making any assumptions about the shape of the underlying feature distribution. A more detailed analysis of the benefits of using EVT is provided in Section V.

[1]https://github.com/RuwanT/EVT-MIL/blob/master/utils/custom_networks.py

The proposed EVT modelling is performed on the intermediate CNN feature space at dense-F layer, i.e. $g_j^{(i)}$ (see Fig 3). To identify the extreme instances in negative bags, we first estimate the negative mean (μ) and the inverse[2] covariance matrix (Σ^{-1}) using the features of correctly classified negative instances. The instance with the maximum Mahalanobis-distance from the mean, in each negative image, is then identified as extreme instances and a Weibull distribution (as it is prescribed by the EVT theorem for the measurement values with finite upper bound) is fitted to those distances using the maximum likelihood estimation method. It should be noted that the CNN feature representations can be sparse and/or may contain highly correlated elements. Furthermore, the EVT fitting procedure assumes that the features of negative instances are closer to negative feature mean (μ) than the positive features. However, a typical neural network classifier, has no constraint to force the feature representations to have the above property. As such, the Mahalanobis-distance is a more appropriate measure of distance in such a feature space compared to the commonly used Euclidean distance.

The Weibull distribution has three parameters: shape (k_w), scale (λ_w) and location (θ_w). The CDF of the above estimated Weibull distribution can then be used to define the positiveness for all instance in positive bags as:

$$P\left(h_j^{(i)} = 1 \mid d_j^{(i)}, k_w, \lambda_w, \theta_w\right) = 1 - e^{-\left(\frac{d_j^{(i)} - \theta_w}{\lambda_w}\right)^{k_w}} \tag{4}$$

where $d_j^{(i)} = \sqrt{\left(g_j^{(i)} - \mu\right) \Sigma^{-1} \left(g_j^{(i)} - \mu\right)}$. Given the precise nature of our positive probability modelling, a simple threshold (we use $T_{evt} = 0.95$ in all of our experiments) or an importance sampling strategy can be applied to infer the values of the hidden variables. This threshold is easy to tune and our experiments, presented in Section V, showed that the final result is not significantly affected by the value of this threshold. To make sure that each positive image has at least one positive instance, we always assign the corresponding hidden variable of the instance with maximum positive probability to one.

As it is likely to have more negative instances than positive ones, we sample 75% of instances from each negative image using the score $\hat{w}_j^{(i)}$ as the probability of selecting a given instance $x_j^{(i)}$. Such sampling strategy (hard negative mining) would increase the likelihood of including an incorrectly classified negative instance in the subsequent training iteration. The overall algorithm for sampling instance labels $h_j^{(i)}$ is given in Algorithm 1. The image level predictions at the test time are determined using equation (2) on CNN output probabilities. The overall CNN training procedure is given in Algorithm 2. At the start of training, the hidden variables ($h_j^{(i)}$) are initialized with the respective bag labels as shown in step 1 of Algorithm 2. Using bag labels as instance labels has been suggested for training MIL classifiers [10], [11], [12] and we will show that the proposed method is capable of significantly improving a classifier performance from this initial point.

[2]The Moore-Penrose pseudo-inverse was used in our implementation as the covariance matrix can be non-invertible.

Algorithm 1 One iteration of inferring h with the EVT.

Input: CNN features at Dense-F ($\{g_j^{(i)}\}$), image labels ($[y^{(i)}]_{i=1}^N$), EVT threshold ($T_{evt} = 0.95$), EVT tail size β.
Output: labelled instances $\{(x_j^{(i)}, h_j^{(i)})\}$

$h_j^{(i)} \leftarrow -1 : \forall\ i$ **and** j
$g_{tn} \leftarrow$ Set of true-negative features.
$\mu \leftarrow Mean(g_{tn})$, $\Sigma^{-1} \leftarrow Cov(g_{tn})^+$
$d_j^{(i)} = sqrt((g_j^{(i)} - \mu)\Sigma^{-1}(g_j^{(i)} - \mu))\ : \forall\ i$ **and** j
$M = \{M^{(i)} \leftarrow \max(d_1^{(i)}, \ldots, d_{L_i}^{(i)}) : \forall\ i$ **if** $y^{(i)} = 0\}$
$[k_w, \lambda_w, \theta_w] \leftarrow$ FitWeibull(M, β)
$w_j^{(i)} \leftarrow 1 - e^{-(\frac{-d_j^{(i)} - \theta_w}{\lambda_w})^{k_w}} : \forall\ i$ **and** j
for $i = 1 \rightarrow N$ where $y^{(i)} = 1$, $j = 1 \rightarrow L_i$ **do**
  if $w_j^{(i)} > T_{evt} : h_j^{(i)} = 1$ **else** $h_j^{(i)} = 0$
  $h_{j*}^{(i)} = 1$ where $j* = argmax(w_1^{(i)}, \ldots, w_{L_i}^{(i)})$
end for
Sample 75% instances from each negative image with acceptance probability $w_j^{(i)}$ and set corresponding $h_j^{(i)} \leftarrow 0$.

Algorithm 2 Training procedure for EVT-MIL.

Inputs: Image instances and bag labels $\{\{x_j^{(i)}\}_{j=1}^{L_i}, y^{(i)}\}_{i=1}^N$.
Output: Trained CNN model

$h_j^{(i)} \leftarrow y^{(i)} : \forall\ i = 1 \rightarrow N$ **and** $j = 1 \rightarrow L_i$
for $i = 1 \rightarrow \#epoch$ **do**
  Train CNN with $\{(x_j^{(i)}, h_j^{(i)})\}$.
  $\{g_j^{(i)}\} \leftarrow$ Extract CNN features at Dense-F.
  $\{(x_j^{(i)}, h_j^{(i)})\} \leftarrow$ Infer instance labels with Algorithm 1.
end for

IV. Experiments

This section describes the experimental set-up used to measure the performances of the proposed EVT-MIL method and compare those with some recently published competitive methods on two different volumetric image classification tasks. The proposed EVT-MIL method was implemented using *keras*[3] library with *TensorFlow* backend. The code for the proposed method is publicly available[4]. The area under the ROC curve (AUC) is used as the main metric in our performance evaluations.

A. 3D retinal OCT image classification

The proposed method was first evaluated on the 3D retinal OCT image classification task in ReTOUCH challenge [15]. OCT is an *in-vivo*, high resolution imaging technology that is capable of capturing a 3D volumetric image of the retinal and the sub-retinal layers as well as their structures. Studies have shown that OCT signals have strong correlation with retinal histology and are extremely useful to diagnose Macular Edema (the swelling of the macula region of the eye, caused by fluid build-ups due to disruptions in blood-retinal barrier [40]) caused by different diseases. In this challenge, the objective is to classify OCT images according to the presence (or absence) of three types of fluids that cause Macular Edema: Intra-retinal fluid (IRF), Sub-retinal fluid (SRF) and Pigment Epithelial Detachment (PED) - although the name does not refer to a fluid, the detachment is marked by existence of excess fluid. An example image patch from the ReTOUCH challenge with annotations indicating the three fluid types is shown in Fig. 4.

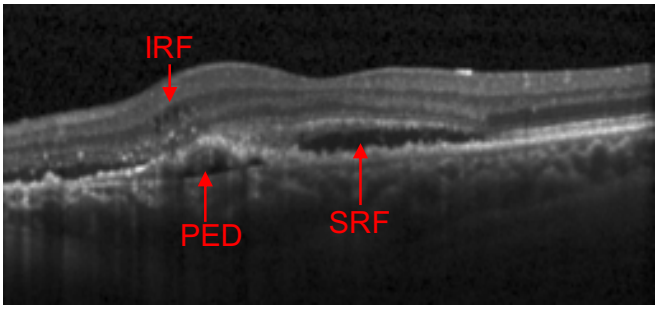


Fig. 4: An example of an image patch (an instance with dimensions $[512 \times 320 \times 1]$) from the ReTOUCH challenge dataset with annotations indicating the existance of three fluid types: Intra-retinal fluid (IRF), Sub-retinal fluid (SRF) and Pigment Epithelial Detachment (PED).

1) Dataset: The ReTOUCH dataset consists of 112 volumetric OCT images of 112 subjects. The images were captured using a variety of devices from three different manufacturers. A detailed description of the dataset is provided in Table I. We have used the train/test split provided by the ReTOUCH organizers in our experiments (60% Train – 40% Test). The ReTOUCH dataset also provides expert segmentations as ground truth (i.e. segmentation masks indicating the presence or absence of the three fluid types for each voxel) for each image in the training set. While we did not use those for training EVT-MIL method, those were used in assessing its performance.

2) Implementation details: For all methods in this section, we extracted image patches with dimensions $512 \times 320 \times 1$ as instances (full b-scan after cropping the boarder as shown in Fig. 4). The cost balancing coefficient λ_1 of the proposed method was set to 0.001. Network parameters were optimized for 100 epoch using "rmsprop" with learning rate 0.001 and decay 1×10^{-8}. The batch size was set to 32. The EVT tail size was set to the minimum of 30 or the number of negative training images.

3) Comparative analysis: We compared the performance of EVT-MIL over three classification tasks (i.e. SRF, IRF, PED classification) with several competing methods, Full-SUP: Fully supervised classifier using expert annotated instance labels derived using the segmentation masks provided with the training data, TOP-K: Our implementation of top k positive instances [9], CNN-Th: Inferring h in the proposed method by thresholding the CNN output instead of using EVT, AttentionDeepMIL: Attention based MIL architecture from [27], MINet: Max-pooling based feature aggregation performed at the last feature layer [32] and Simple-MIL: Using bag labels as instance labels to train CNN. To make the comparisons fair we used the same architecture and hyperparameters during the

[3]https://github.com/keras-team/keras.git
[4]https://github.com/RuwanT/EVT-MIL.git

TABLE I: Details of the ReTOUCH OCT image dataset.

	Spectralis	Cirrus	Topcon
Number of Images	38	38	36
Number of Positives (IRF/SRF/PED)	32/22/19	30/19/18	26/20/16
B-scans Size	512 × 496	512 × 1024	512 × 885/650
Number of B-scans	49	128	128
Axial resolution	3.9μm	2.0μm	2.6/3.5μm

training of each method above.

Fig 5 shows the mean and standard deviation of the test-set AUC values in image level fluid type classification over 25 repeated runs of each method. As expected, the figure shows that the Full-SUP method has been able to achieve the best AUC values across all three fluid types. The proposed method, trained without any instance level supervision, has also achieved comparable AUC values in classifying SRF and PED, indicating that it is possible to learn a good image level predictor in a weakly supervised manner.

The p-values from a two-sided Wilcoxon signed-rank test between the pairs of average classification errors (EVT-MIL and the competing method) for each image in the test set are provided in Table II. The average classification error for each image for a given method is the error averaged over the 25 repeated runs of that method where the treshold of each classifier is set in such a way that it yields a true-positive-rate of 80%. Symbol (R) in this table indicates that the null hypothesis (the difference between the medians of two methods being zero) can be rejected with 95% confidence. The results also show that the AUC archived by the proposed method is significantly better than that of TOP-K in SRF and PED. This demonstrates the effectiveness of using an adaptive number of positive instances in the proposed method compared to using fixed thresholds. The methods Simple-MIL and CNN-Th also generate poor results, compared to the proposed method, in SRF and PED classification.

The results of the proposed method have the largest deviation form Full-SUP in IRF fluid type. This can be attributed to the limited number of negative samples present in IRF classification task compared to the other two fluid types. As shown in Table I, the number of negative bags for IRF task is only 24 compared to 51 for SRF and 59 for PED. The empirical results shown in Fig 14 also indicate that limiting the number of negative bags at training time would diminish the performance of EVT-MIL.

The top performing method in the Re-TOUCH challenge (i.e. SFU [15]) produces AUC scores of 1.0 for all three fluid types. However, the above method uses expert annotated fluid segmentation masks for training and uses an off-the-shelf layer segmentation method. Considering that our proposed method only uses image level labels, these results are not directly comparable.

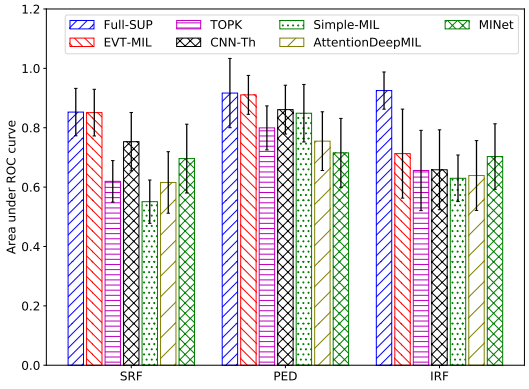
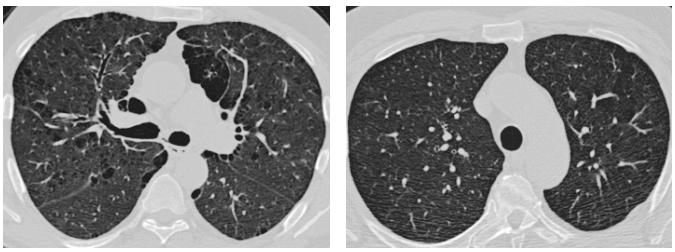


Fig. 5: The test-set AUC values in image level fluid type classification. The plots show the mean and the standard deviation of AUC values over 25 repeated runs for each method. Full-SUP: Fully supervised classifier trained using expert annotated instance labels, CNN-Th: Instance labels updated by thresholding CNN output at each epoch, Simple-MIL: Bag labels used as instance labels, TOP-K [9], AttentionDeepMIL [27], MINet [32] and EVT-MIL: Proposed method.

B. *Emphysema detection in pulmonary 3D-CT images*

In this section, we evaluate the proposed method for Emphysema detection in low-dose CT images. Emphysema is a lung pathology characterized by the destruction of lung tissue and enlargement of airspaces in a lung. It is part of chronic obstructive pulmonary disease (COPD), which is a leading cause of mortality and morbidity worldwide [41]. Visual assessment of emphysema based on CT is presumed to be more sensitive to emphysema than CT densitometry [42]. Examples of lung regions with and without emphysema are shown in Fig 6.



(a) (b)

Fig. 6: Axial slices of two 3D-CT images with (left) and without (right) emphysema.

The data used in this section is collected during the Danish Lung Cancer Screening Trial (DLCST) [43]. The scan parameters as described by [43] are: "All CT scans of the study were performed on a MDCT scanner (16 rows Philips Mx 8000, Philips Medical Systems, Eindhoven, The Netherlands). Scans were performed supine after full inspiration with caudocranial scan direction including the entire ribcage and upper abdomen

TABLE II: The p-values from two sided Wilcoxon signed-rank test between EVT-MIL and each competing method on the three fluid classification tasks. Symbol (R) indicates that the null hypothesis (the difference between the means of two methods being zero) can be rejected with 95% confidence.

	Full-SUP	TOP-K	CNN-Th	Simple-MIL	AttentionDeepMIL	MINet
SRF	4.81E-01	3.83E-02 (R)	9.24E-02	3.49E-05 (R)	7.22E-05 (R)	1.17E-02 (R)
PED	9.87E-01	1.66E-02 (R)	1.19E-02 (R)	2.89E-05 (R)	4.54E-03 (R)	4.43E-03 (R)
IRF	4.46E-04 (R)	5.65E-03 (R)	1.83E-02 (R)	5.35E-06 (R)	3.94E-02 (R)	1.13E-02 (R)

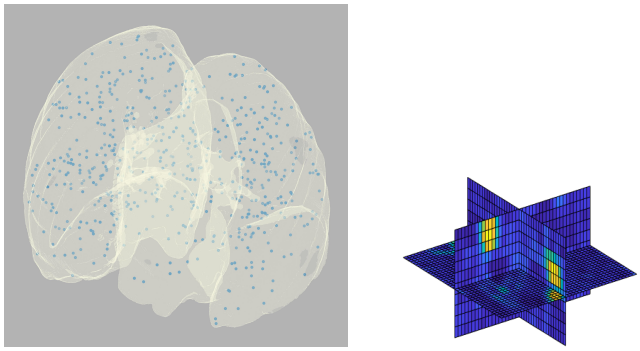
with a low dose technique, 120kV and 40 mAs. Scans were performed with spiral data acquisition with the following acquisition parameters: Section collimation 16 $\times$ 0.75 mm, pitch 1.5, rotation time 0.5 second". The images were reconstructed with slice thickness of 1mm, using a hard convolutional kernel.

1) Dataset: We sampled 200 scans (100 with emphysema and 100 without emphysema) from the DLCST dataset and used them in our evaluations. The annotations (has emphysema vs. no emphysema) for each image were derived from the visual assessments described in Wille *et al.* [42], where each lung image was divided into six regions (the regions were defined as above carina, between carina and lower pulmonary vein, and below lower pulmonary vein) and two experts assessed the extent of emphysema in each of those regions. The extent was assessed as a categorical grade ranging from 0 to 5 corresponding to 0%, 1-5%, 6-25%, 26-50%, 51-75% and 76-100% of emphysema tissue respectively. In our experiments, we used the annotations from one expert to train and test the models and use the annotations from the other expert to calculate the inter-observer agreement. Table III shows the rater agreement on region level emphysema presence in the sampled dataset. The data show that the rater agreement is higher in the upper regions of the lungs compared to lower regions.

The sampled dataset was divided into four fixed folds and four-fold cross validation was used in our experiments. The cross validation splits were selected such that they include the similar proportion of positives and negatives.

2) Implementation details: 600 possibly overlapping patches from each image (100 patches from each region) with dimensions $41 \times 41 \times 21$ voxels were used as instances for the proposed method. The distribution of instances and a randomly selected instance for an example 3D-CT image is visualized in Fig. 7. The cost balancing coefficients λ_1 of the proposed method was set to 1×10^{-6}. Network parameters were optimized for 60 epoch using "rmsprop" with learning rate 0.001 and decay 1×10^{-8}. The batch size was set to 16. The EVT tail size was set to the minimum of 50 or the number of negative training images.

3) Comparative analysis: In this section, we compare the performance of EVT-MIL with three published methods on the above dataset: 1) Ørting *et al.* [12]: Simple-MIL based approach that uses hand-crafted features (i.e. equalized histograms of multi-scale filter representations) within a logistic regression framework. 2) GAP-Net [44]: Deep learning based model that takes a whole lung region as input and predicts the severity score. This method uses the region-level emphysema precence annotations to generate the training signal. 3) Prop-Net [44]: Similar to GAPNet but uses an enhanced architecture specialized for label proportion learning with an improved loss function. It should be noted that methods EVT-MIL and [12] uses image level labels where as the other two methods, i.e. Prop-Net and GAP-Net, use region level labels.



(a) (b)

Fig. 7: (a) 3D visualization of the distribution of instance locations within a 3D-CT image. Each blue dot indicate a center of an instance. (b) A slice plot of an instance with dimensions $41 \times 41 \times 21$ voxel.

The cross-validation results of the proposed method together with the three competing methods are shown in Table IV. The results show that the proposed method has been able to achieve AUC values that are significantly better than those of [12]. This suggests that the proposed method has been able to learn features that are superior to the hand-crafted features used by [12]. The results also show that EVT-MIL trained only with image level labels has been able to achieve similar accuracy to Prop-Net, which uses significantly more information during training (i.e. region level emphysema severity scores). Prop-Net can be viewed as a fully-supervised deep learning method, when it comes to region level emphysema prediction, as it uses a whole image region as input to the network together with the corresponding emphysema presence label.

It should be noted that using bag labels as instance labels to train (Simple-MIL) the deep network failed to converge for this dataset, hence the results were not included (the features collapsed to a single point in the feature space).

Fig. 8 shows the ROC curves for each competing method on the emphysema detection task, which again shows that the proposed EVT-MIL method has been able to achieve high accuracy comparable to Prop-Net. The figure also shows that the results of the proposed method is in-line with the inter-observer agreement (annotations of two different experts).

C. 2D Histopathology image classification

In this section, we evaluate the proposed method on automatic detection of cancerous regions in 2D histopathology

TABLE III: Rater agreement on region level emphysema presence in the dataset. Different regions of lung are indicated by: ERL=right-lower, ERM=right-middle, ERU=right-upper, ELL=left-lower, ELM=left-middle, ELU=left-upper.

	ERL	ERM	ERU	ELL	ELM	ELU	Scan
Absent	92.72	96.29	97.24	90.64	95.74	96.69	96.15
Present	68.57	61.53	74.72	68.96	62.71	73.41	71.88

TABLE IV: Four-fold cross-validation results (AUC) for emphysema detection using low-dose CT images. ERL=right-lower, ERM=right-middle, ERU=right-upper, ELL=left-lower, ELM=left-middle, ELU=left-upper region of lung.

		ERL	ERM	ERU	ELL	ELM	ELU	Scan
Ørting et al. [12]	Mean	0.75	0.83	0.88	0.74	0.76	0.85	0.84
	Std	0.09	0.02	0.07	0.13	0.03	0.07	0.07
GAP-Net [44]	Mean	**0.92**	0.93	0.94	0.93	0.91	0.96	0.93
	Std	0.03	0.03	0.02	0.06	0.03	0.00	0.02
Prop-Net [44]	Mean	0.90	**0.95**	**0.96**	**0.96**	**0.93**	**0.98**	**0.95**
	Std	0.06	0.02	0.02	0.02	0.03	0.02	0.04
EVT-MIL	Mean	0.90	**0.95**	**0.96**	0.95	**0.93**	0.96	**0.95**
	Std	0.06	0.01	0.01	0.05	0.01	0.01	0.06

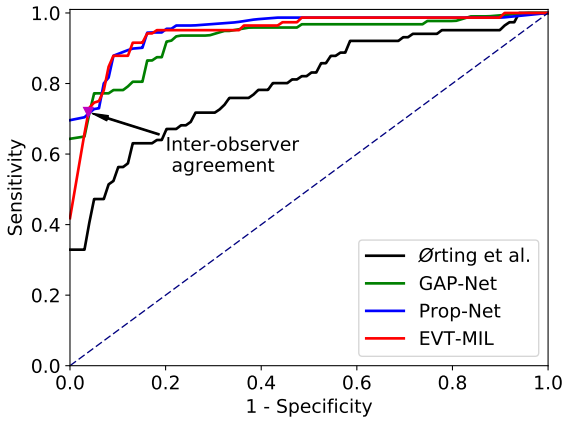
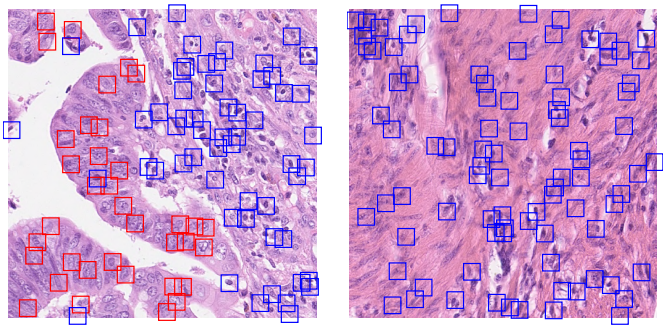


Fig. 8: ROC curves of different emphysema detection methods. The figure shows that the proposed EVT-MIL method has been able to achieve high accuracy. The plot also show the inter-observer agreement and indicates that the EVT-MIL results are comparable to the inter-observer agreement.



(a) Positive (b) Negative

Fig. 9: A positive and a negative bag from the whole-slide histopathology image dataset. The rectangle annotations indicate the instance (with dimensions $[27 \times 27]$) extracted using manually annotated cell nuclei (for clarity, only a subset of instances are visualized). The color of the rectangle indicate instance label (red for positive and blue for negative).

images of colorectal adenocarcinomas. Although the focus of this paper is on volumetric images, histopathology images are often very large (high resolution) and share some of the issues discussed in this paper related to training deep CNN models with volumetric images. Furthermore, 2D image classification experiments simplify the visualization process and help with understanding EVT-MIL while enable the comparisons with general MIL methods.

1) Dataset: The used dataset comprises of 100 hematoxylin and eosin stained whole-slide histopathelogy images of colorectal adenocarcinomas. A detailed description of the dataset can be found in [26]. The locations of 24,444 cell nuclei, across all the images in the dataset, has been manually annotated together with their associated class label (i.e. epithelial, inflammatory, fibroblast and miscellaneous). A bag used in our experiments is composed of 27×27 2D image patches centred on manually annotated cell nuclei. A bag is labelled positive if it contains one or more patches from the epithelial class and zero otherwise. In total there are 51 positive bags. Two example histopathelogy images from the above dataset together with a subset of extracted instances are shown in Fig. 9.

2) Implementation details: We used the classification CNN architecture proposed in [26] as the base network. The network parameters were optimized for 50 epoch using "Adam" optimizer with the same hyper-parameters used in [27].

3) Comparative analysis: To analyze the performance of the proposed method, ten-fold cross-validation results for EVT-MIL on histopathology classification dataset is reported in Table V together with the results reported in [27]. In the latter work, authors proposed an end-to-end trained MIL network that employed several instance aggregation regimes. Their main contribution was a weighted averaging of instance level outputs using a permutation invariant aggregation operator that corresponds to attention mechanism (Attention, Gated-Attention). Furthermore, they also used non learned max aggregation, which either operates on instance predictions (Instance-max) or on feature embeddings (Embedding-max). Embedding-max, in which feature aggregation is performed at the last feature layer, is almost equivalent to MI-Net with

max-pooling [32].

Table V shows that the EVT-MIL, which employs non learned instance aggregation method, has achieved higher AUC values compared to other similar methods (Instance-max, Embedding-max). Furthermore, EVT-MIL has achieved similar performance to learning based instance aggregation methods: Attention and Gated-Attention. Several examples of successful and failed cases for EVT-MIL is visualized in Fig 10.

V. Discussion

The paper presents a novel multiple instance learning based method for training a deep neural network for volumetric image classification that can be used when it is not possible to train a deep network with the whole image due to GPU memory restrictions. The proposed method can adaptively select positive instances from positive bags at training time. The method is aimed at situations where the number of positive instances in a bag varies in a broad range (hence defining a fixed threshold indicating the expected number of positive instances per bag would not be accurate) and where inclusion of all the positive instances in training is important (due to limited number of training bags) for learning an accurate model.

In the following sections, we discuss the key properties of the proposed method. If not stated otherwise, the results from the emphysema detection task is used in the analyses (due to relative large and balanced dataset).

A. Why use EVT modelling?

A main advantage of the proposed method is the use of extreme value theorem to model the deviation of true negative features from their mean. To demonstrate the advantage of using EVT modelling, we extracted the features for all the instances in the training set at a specified epoch during training and computed the distances $d_j^{(i)}$ as defined in Algorithm 1. The normalized histogram of those distances for all instances in positive bags and true negative instances are plotted in Fig. 11. The figure shows that the distribution of distances of true-negative instances has an arbitrary shape and assuming a specific underlying shape to model this distribution would not be reasonable (e.g. using a Gaussian to model the distribution of mean-deviation in true-negative instances). EVT does not assume any underlying distribution and hence is an ideal tool in such applications.

We have also plotted the Weibull CDF estimated at the same point in Fig. 11. The plot shows that the EVT estimation has been able to effectively differentiate the main clusters in the positive bags. The figure also demonstrates that the EVT modelling has resulted in a sharp CDF which helps reduce the sensitivity to the chosen EVT threshold. The results presented in Fig. 12, which shows the variation of model accuracy with the EVT threshold, also supports the argument that changing the EVT threshold in a wide range does not significantly vary the performance of the proposed method.

To gain a better understanding of the feature representations learnde by the proposed EVT-MIL method, we have plotted the t-SNE plot of instance features in Fig. 13. The t-SNE plots show that features of instances that belong to the negative bags are clustered around the negative mean, whereas the features of instances from positive bags show a distinct cluster in addition to the main cluster which overlaps the instances from negative bags.The plot shows that the proposed method has learned a feature representation that is discriminative at instance level.

B. Training dataset size vs accuracy.

To measure the effect of training dataset size on the accuracy of the resulting model, we removed different portions of images from each training fold (randomly sampled) and calculated the cross-validation values. Those results are shown in Fig. 14. Test sets for each fold remain unchanged. The figure shows that the proposed method has been able to learn accurate models when the dataset size is between 100% (200 images) to 50% of the original dataset size. The accuracy of the learned models starts to degrade when the training dataset is reduced to a lesser than 50% of the original dataset size. At 50% each fold contains around 75 images (approximately 35 images from each class). The results also demonstrate that the proposed method is capable of learning accurate models even when the size of the training dataset is fairly limited.

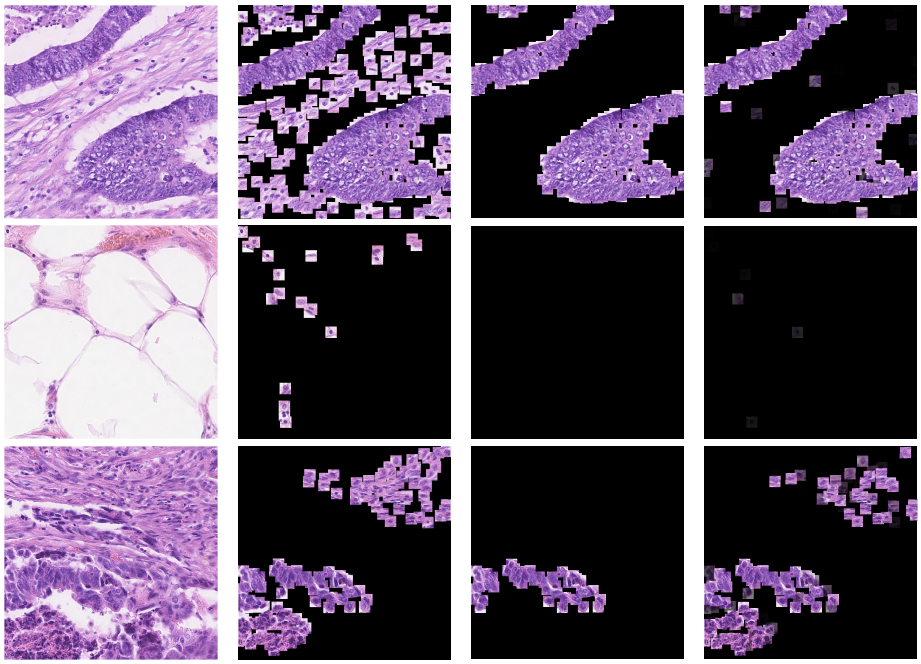
Points M1 and M2 on Fig. 14 show the mean AUC values for models trained with 65% data where either negative (M1) or positive (M2) images were removed from the dataset. The results indicate that the learning process is more sensitive to loss of negative training data than to the loss of positives. This can be viewed as a limitation of the proposed method and can perhaps be explained by the fact that we only use negatives to estimate the boundary between positive and negative instances. The extreme value theorem relies on the number of samples to be large. However, in the above scenario, the number of samples (the number of negative images in the training dataset) is limited leading to a sub-standard extreme value estimation. The above observation points out that it is appropriate to use the method in circumstances when the number of negative bags is relatively large. In many practical applications it may be easier to collect large number of negative images compared to images containing specific abnormality.

C. Effectiveness of selecting positive instances from positive bags

All weakly-supervised or MIL classifiers learn to label the whole image. Predicting the class label of the entire image alone might not be sufficient for many clinical applications where it is important to identify regions in the image that lead to the overall decision. A subset of MIL classifiers that falls under the instance-space paradigm described in Section II has the ability to classify individual patches (or instances), thus these methods can indicate regions with signs of pathology. The proposed method falls into the latter category, where we can use the learned EVT model to generate the positiveness probability for all the instances of a test image. Fig. 10 shows several example cases from the 2D histopathology image classification task. Column (d) visualizes the instance level positiveness probability (generated by EVT-MIL) by scaling

TABLE V: Ten-fold crssvalidation results (AUC) on histopathology classification dataset [26]. Experiments were run 5 times and the average and standard deviation of mean is reported. The results for the competing methods are taken from [27].

	Instance-max	Embedding+max	Attention	Gated-Attention	EVT-MIL
AUC	0.914 ± 0.010	0.918 ± 0.010	0.968 ± 0.009	0.968 ± 0.010	0.966 ± 0.008



(a) (b) (c) (d)

Fig. 10: Depiction of successful or failed cases of the 2D Histopathology image classification task. Columns represent: (a) Histopathology image, (b) All extracted instances of size 27×27, (c) Ground truth: Patches belong to class epithelial (d) EVT-MIL prediction: The intensity of each patch in column b is scaled with the positiveness probability predicted by EVT-MIL. Rows one and two show two cases where EVT-MIL has been successful. The third row shows a case where the predicted bag label is correct but the instance level predictions of EVT-MIL contain false positives.

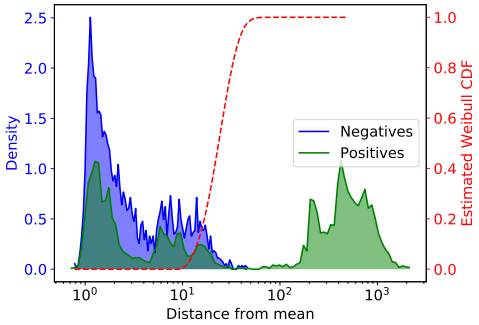


Fig. 11: Normalized histogram of distance to the mean in the feature space for positive and negative instances together with the estimated Weibull CDF (red dash line) at a selected point in the training process.

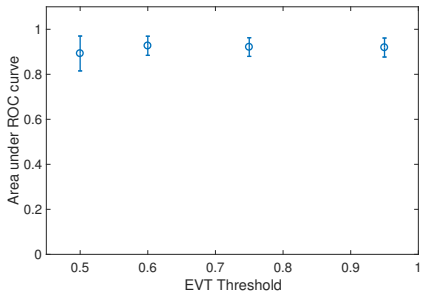
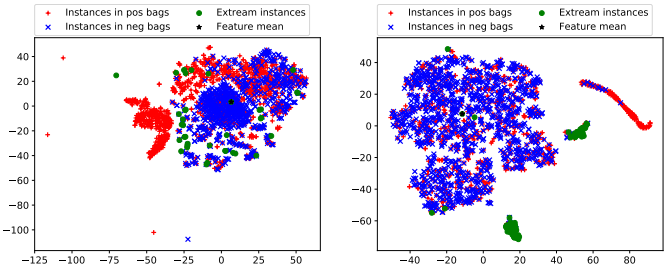


Fig. 12: Variation of cross-validation accuracy with values of EVT threshold in the range of $[0.5 - 1.0]$ for the emphysema detection task.

the intensity of each patch in the image by the positiveness probability. The first two rows of the figure show that the positiveness probability from EVT-MIL can be used to identify the positive lesions in an image correctly. Row three however shows that in few cases EVT-MIL predicts inaccurate positiveness probabilities while predicting the correct bag level label. Learning sub-optimal decision functions is a known limitation of MIL or weakly supervised learning methods and detailed analysis of this shortcoming can be found in [45], [33].

Fig. 15 shows the evolution of the positive instance selection process during training. The number of positive instances selected by EVT-MIL method from each positive image gradually increases during training. This would in turn enable more



(a) (b)

Fig. 13: t-SNE embedding of the learned feature representation at Dense-F layer for (a) 2D Histopathology image classification, (b) Emphysema detection in 3D-CT. The instances from negative bags are represented by blue crosses whereas, the instances from positive bags are represented by red plus signs. The green dots represent extreme instances used in learning the Weibull parameters and the black star represents the negative feature mean. The plot shows that the proposed method has learned a feature representation that is discriminative at an instance level.

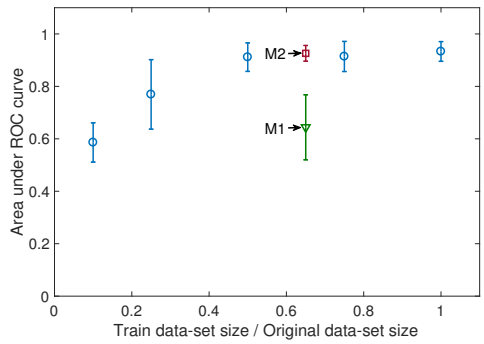


Fig. 14: Variation of cross-validation accuracy (measured with AUC) with the size of dataset. The size of the dataset is represented as a proportion of the original dataset size (200 scans). M1, M2 show the mean AUC values for models trained with 65% data where only negative images (M1) or positive images (M2) were removed from the dataset.

accurate models to be learned. Furthermore, the number of selected positive instances show positive correlation with the emphysema grade of the scan indicating the effectiveness of the positive instance selection process of the proposed method.

VI. Conclusion

The paper presents a new method for training a deep neural network for volumetric image classification. In the proposed method, the original problem is posed as a multi-instance classification problem and the extreme value theorem is used to infer the instance level labels given the current state of intermediate CNN features. The experimental results on fluid type classification in OCT, Emphysema detection in low-dose CT images and automatic detection of cancerous regions in 2D histopathology images show that using the EVT to infer the instance labels significantly improves over using a threshold scheme as performed in state-of-the-art methods.

Acknowledgment

This research was funded partially by the Australian Research Council Linkage Project grant (LP160100662), Netherlands Organization for Scientific Research (NWO) and Research Fund Denmark (DFF).

References

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Snchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, no. Supplement C, pp. 60 – 88, 2017.

[2] H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, May 2016.

[3] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, W. Xie, G. L. Rosen, B. J. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A. E. Carpenter, A. Shrikumar, J. Xu, E. M. Cofer, C. A. Lavender, S. C. Turaga, A. M. Alexandari, Z. Lu, D. J. Harris, D. DeCaprio, Y. Qi, A. Kundaje, Y. Peng, L. K. Wiley, M. H. S. Segler, S. M. Boca, S. J. Swamidass, A. Huang, A. Gitter, and C. S. Greene, "Opportunities and obstacles for deep learning in biology and medicine," *Journal of The Royal Society Interface*, vol. 15, no. 141, 2018.

[4] H. R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. M. Summers, "Improving computer-aided detection using convolutional neural networks and random view aggregation," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1170–1181, May 2016.

[5] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1207–1216, May 2016.

[6] "Solution of grt123 team," https://github.com/lfz/DSB2017/blob/master/solution-grt123-team.pdf.

[7] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pèrez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1, pp. 31 – 71, 1997.

[8] G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard, "Multiple-instance learning for medical image and video analysis," *IEEE Reviews in Biomedical Engineering*, vol. PP, no. 99, pp. 1–1, 2017.

[9] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 603–611.

[10] V. Cheplygina, L. Sørensen, D. M. Tax, J. H. Pedersen, M. Loog, and M. de Bruijne, "Classification of COPD with multiple instance learning," in *22nd International Conference on Pattern Recognition (ICPR)*. IEEE, 2014, pp. 1508–1513.

[11] L. Sorensen, M. Nielsen, P. Lo, H. Ashraf, J. H. Pedersen, and M. De Bruijne, "Texture-based analysis of COPD: a data-driven approach," *IEEE transactions on medical imaging*, vol. 31, no. 1, pp. 70–78, 2012.

[12] S. N. Ørting, J. Petersen, L. H. Thomsen, M. M. W. Wille, and M. de Bruijne, "Detecting emphysema with multiple instance learning," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, April 2018, pp. 510–513.

[13] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in neural information processing systems*, 2003, pp. 577–584.

[14] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and J. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2424–2433.

[15] H. Bogunovi, F. Venhuizen, S. Klimscha, S. Apostolopoulos, A. Bab-Hadiashar, U. Bagci, M. F. Beg, L. Bekalo, Q. Chen, C. Ciller, K. Gopinath, A. K. Gostar, K. Jeon, Z. Ji, S. H. Kang, D. D. Koozekanani, D. Lu, D. Morley, K. K. Parhi, H. S. Park, A. Rashno, M. Sarunic, S. Shaikh, J. Sivaswamy, R. Tennakoon, S. Yadav, S. De Zanet, S. M. Waldstein, B. S. Gerendas, C. Klaver, C. I. Snchez, and U. Schmidt-Erfurth, "RETOUCH – The retinal OCT fluid detection and

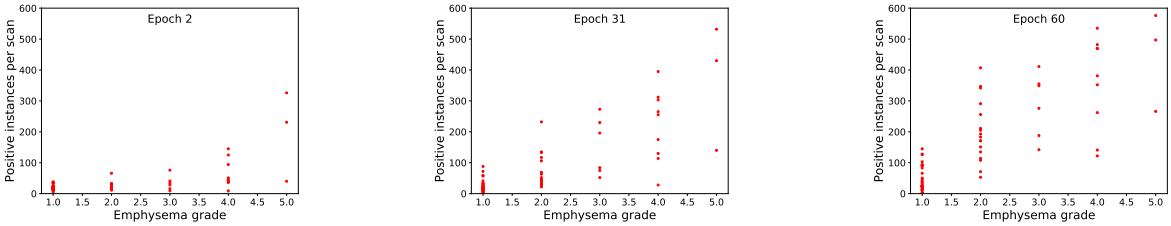


Fig. 15: Number of positive instances selected at three different epochs during the training process of EVT-MIL are plotted against the emphysema grade of the scan. The selected number of positive instances increases during the training cycle and EVT-MIL has correctly identified more positive instances from scans with higher emphysema grade.

segmentation benchmark and challenge," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2019.

[16] W. J. Scheirer, A. Rocha, R. J. Micheals, and T. E. Boult, "Meta-recognition: The theory and practice of recognition score analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 8, pp. 1689–1695, 2011.

[17] W. J. Scheirer, A. Rocha, A. Sapkota, and T. E. Boult, "Towards open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 35, July 2013.

[18] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1563–1572.

[19] R. Tennakoon, A. K. Gostar, R. Hoseinnezhad, M. de Bruijne, and A. Bab-Hadiashar, "Deep multi-instance volumetric image classification with extreme value distributions," in *Asian Conference on Computer Vision (ACCV)*. Cham: Springer International Publishing, 2018.

[20] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. van Ginneken, "Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1160–1169, 2016.

[21] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[22] Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V. C. Mok, L. Shi, and P.-A. Heng, "Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1182–1195, 2016.

[23] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani, "Deep features learning for medical image analysis with convolutional autoencoder neural network," *IEEE Transactions on Big Data*, vol. PP, no. 99, pp. 1–1, 2017.

[24] Z. Yan, Y. Zhan, Z. Peng, S. Liao, Y. Shinagawa, S. Zhang, D. N. Metaxas, and X. S. Zhou, "Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1332–1343, 2016.

[25] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial Intelligence*, vol. 201, pp. 81 – 105, 2013.

[26] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. Snead, I. A. Cree, and N. M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1196–1206, 2016.

[27] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International Conference on Machine Learning*, 2018, pp. 2132–2141.

[28] G. Quellec, M. Lamard, M. D. Abrmoff, E. Decencire, B. Lay, A. Erginay, B. Cochener, and G. Cazuguel, "A multiple-instance learning framework for diabetic retinopathy screening," *Medical Image Analysis*, vol. 16, no. 6, pp. 1228 – 1240, 2012.

[29] G. Quellec, M. Lamard, A. Erginay, A. Chabouis, P. Massin, B. Cochener, and G. Cazuguel, "Automatic detection of referral patients due to retinal pathologies through data mining," *Medical Image Analysis*, vol. 29, pp. 47 – 64, 2016.

[30] J. Wang and J.-D. Zucker, "Solving multiple-instance problem: a lazy learning approach," in *International Conference on Machine Learning*. Morgan Kaufmann Publishers, 2000, pp. 1119–1126.

[31] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola, "Multi-instance kernels," in *Proceedings of the Nineteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 2002, pp. 179–186.

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

[33] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329 – 353, 2018.

[34] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical Image Analysis*, vol. 54, pp. 280 – 296, 2019.

[35] S. Coles, *An introduction to statistical modeling of extreme values*. Springer, 2001, vol. 208.

[36] S. M. Berman, "Limiting distribution of the maximum term in sequences of dependent random variables," *The Annals of mathematical statistics*, vol. 33, no. 3, pp. 894–908, 1962.

[37] A. Ferreira, L. De Haan *et al.*, "On the block maxima method in extreme value theory: Pwm estimators," *The Annals of statistics*, vol. 43, no. 1, pp. 276–298, 2015.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[39] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and $<$ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[40] M. F. Marmor, "Mechanisms of fluid accumulation in retinal edema," in *Macular Edema*. Springer, 2000, pp. 35–45.

[41] R. A. Pauwels and K. F. Rabe, "Burden and clinical features of chronic obstructive pulmonary disease (COPD)," *The Lancet*, vol. 364, no. 9434, pp. 613 – 620, 2004.

[42] M. M. W. Wille, L. H. Thomsen, A. Dirksen, J. Petersen, J. H. Pedersen, and S. B. Shaker, "Emphysema progression is visually detectable in low-dose CT in continuous but not in former smokers," *European Radiology*, vol. 24, no. 11, pp. 2692–2699, Nov 2014.

[43] J. H. Pedersen, H. Ashraf, A. Dirksen, K. Bach, H. Hansen, P. Toennesen, H. Thorsen, J. Brodersen, B. G. Skov, M. Døssing, J. Mortensen, K. Richter, P. Clementsen, and N. Seersholm, "The Danish randomized lung cancer CT screening trial overall design and results of the prevalence round," *Journal of Thoracic Oncology*, vol. 4, no. 5, pp. 608 – 614, 2009.

[44] G. Bortsova, F. Dubost, S. Ørting, I. Katramados, L. Hogeweg, L. Thomsen, M. Wille, and M. de Bruijne, "Deep learning from label proportions for emphysema quantification," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham: Springer International Publishing, 2018, pp. 768–776.

[45] V. Cheplygina, L. Sørensen, D. M. J. Tax, M. de Bruijne, and M. Loog, "Label stability in multiple instance learning," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 539–546.