# Detecting Multiple Myeloma via Generalized Multiple-Instance Learning

Jan Hering[a], Jan Kybic[a], and Lukáš Lambert[b]

[a]Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic
[b]Department of Radiology, First Faculty of Medicine, Charles University in Prague, Czech Republic

## ABSTRACT

We address the task of automatic detection of lesions caused by multiple myeloma (MM) in femurs or other long bones from CT data. Such detection is already an important part of the multiple myeloma diagnosis and staging. However, it is so far performed mostly manually, which is very time consuming. We formulate the detection as a multiple instance learning (MIL) problem, where instances are grouped into bags and only bag labels are available. In our case, instances are regions in the image and bags correspond to images. This has the advantage of requiring only subject-level annotation (ground truth), which is much easier to get than voxel-level manual segmentation. We consider a generalization of the standard MIL formulation where we introduce a threshold on the number of required positive instances in positive bags. This corresponds better to the classification procedure used by the radiology experts and is more robust with respect to false positive instances. We extend several existing MIL algorithms to solve the generalized case by estimating the threshold during learning. We compare the proposed methods with the baseline method on a dataset of 220 subjects. We show that the generalized MIL formulation outperforms standard MIL methods for this task. For the task of distinguishing between healthy controls and MM patients with infiltrations, our best method makes almost no mistakes with a mean AUC of 0.982 and $F_1 = 0.965$. We outperform the baseline method significantly in all conducted experiments.

**Keywords:** multiple-instance learning, multiples myeloma, classification, CAD

## 1. PURPOSE

Multiple Myeloma (MM) is a heterogeneous, malignant clonal plasma cell disorder that results in infiltrations of bone marrow and osteolytic lesions of the skeleton. Diagnosis and treatment planning of multiple myeloma are based on blood and urine tests, bone marrow biopsy as well as imaging (radiography, CT, PET-CT and MRI).[1,2] The imaging information plays an important role in the treatment decision making process,[3] but the evaluation is still mostly performed manually by the radiologists, which is very time consuming. Here, we shall consider the task of detecting MM in femurs (Figure 1(A, B)), with applicability to other long bones. An automatic system would be highly beneficial, especially in the screening context. To our best knowledge, there is only one fully automated method available.[4] It uses histograms to build a density model of normal tissue and detects lesions as outliers to this model. Training data annotation is only available at the image or subject level but not for individual voxels. For this reason, abnormal tissue is not modeled by the method.[4] The lack of voxel-level annotations also limits the use of traditional machine-learning (ML) approaches.[5] Global annotations at subject level like the disease stage are available at low efforts on the manual task, but more challenging to the classification approach. Here we approach the task of MM detection in femurs with only image-level annotations from another angle, by formulating it as a generalization of the Multiple-instance learning (MIL) problem. MIL is a paradigm, where instances (e.g. voxels or image regions) are grouped into bags (e.g. images) and only bag labels are available. In *standard* MIL, a bag is considered positive iff at least one of its instances is positive. MIL was successfully applied to automatic detection tasks in retinopathy,[6] mammography[7] or tuberculosis.[8]

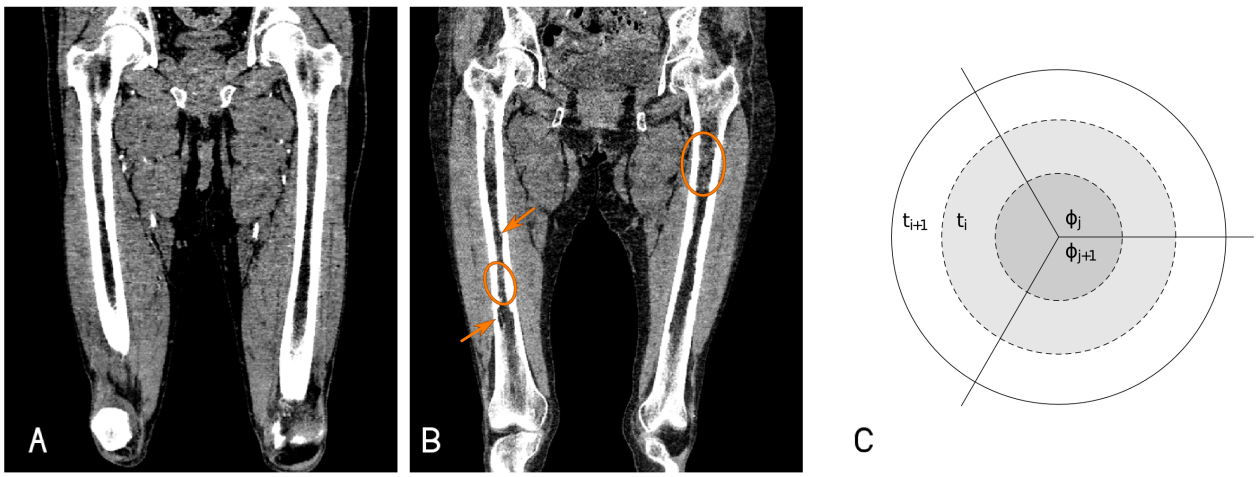


Figure 1: Example low-dose CT femur images: (A) A complete image of a healthy control group, (B) A MM-diseased subjects showing infiltrated areas (marked by ellipses) as well as affected bone tissue (arrows)). (C) Illustration of splitting the bone marrow volume into segments.

## 2. METHODS

### 2.1 Feature extraction

First, the bone marrow region is identified by thresholding (to detect and mask out the bone) and morphological operations.[4] The femurs are aligned with the $z$ axis. The resulting region is divided into 3D angular regions of interest (ROIs) in cylindrical coordinates by uniformly choosing $N_z$ intervals along the $z-$axis, $N_t$ intervals of the radius and $N_\phi$ intervals of the angle coordinate (see Figure 1(C) for illustration). From each ROI a vector of intensity features is extracted — the mean $\mu$, the standard deviation $\sigma$ and a histogram $\{h_i\}$ with $N_h$ bins. Together with the cylindrical coordinates $(l, t, \phi)$, the concatenated vector

$$\mathbf{x}_i = [l, t, \phi, \mu, \sigma, h_1, \ldots, h_{N_h}]$$

represents one ROI, which is regarded as one MIL instance. All ROIs from both femurs of a subject together form a bag $B_I \in \{-1, +1\}$ with a bag label $Y_I \in \{-1, +1\}$, $-1$ meaning negative or healthy.

### 2.2 Multiple instance learning algorithms

Multiple-instance learning (MIL) methods operate on a set of *bags* with labels $(B_I, Y_I)$, $I \in \mathcal{I}$, $Y_I \in \{-1, +1\}$. Each bag consists of a set of *instances* $B_I = \{x_1, \ldots, x_{N_I}\}$, associated with (hidden) labels $y_i \in \{-1, +1\}$. The central principle of MIL is the *positive identifiability* constraint, which defines a bag as positive if at least one of its instances is identified as positive ($y_i = 1$)

$$\hat{Y_I} = \max_{i \in I} \hat{y_i}, \tag{1}$$

and the complementary, *negative exclusion* constraint, expecting only instances with negative labels in negative bags.

The standard MIL algorithms are often formulated as iterative,[9–11] starting with an initial guess of the instance labels and alternating between optimizing the instance classifier from the estimated instance labels and correcting instance labels by combining classifier output and bag labels.

Andrews *et al.*[9] proposed two heuristics to generalize the soft-margin Support Vector Machine (SVM) classifier to solve the MIL task. **mi-SVM** first estimates instance labels given bag constraints and maximizes the margin between instances. The heuristics starts with all instance labels set to the corresponding bag label and iterates until convergence, i.e. until none of the (hidden) instance labels changes. The heuristics verifies in each

iteration if at least one positive instance is present in each positive bag, and sets the imputed label for the instance with maximal decision function value $y(i^*) = 1$ if the constraint is violated.

**MI-SVM** represents each positive bag by its 'top' instance regarding the decision function value (the *witness*) and disregards the remaining instances – the underlying SVM is then iteratively re-trained until the set of the 'top' instances remains unchanged.

Leistner *et al.*[10] addressed the MIL-classification via Random Forests (RF) and deterministic annealing strategy and 'soft' instance labels modeled by the annealing's probabilistic distribution $p$. The minimization of $p$ can be performed either exactly (**MIL-RF**), or approximately (**MIL-ARF**) by Monte-Carlo sampling.[10] The *positive identifiability* constraint is enforced in the same manner as for **mi-SVM**.

### 2.3 Generalized MIL

MIL classification can be seen as a two-level task of identifying the instance labels and aggregating them into bag labels. In standard MIL, the aggregation function is simply

$$\hat{Y}_I = \max_{i \in I} \hat{y}_i,$$

considering a bag to be positive iff there is at least one positive instance label.

A straightforward generalization is to require the number of positive instances to exceed a given threshold $\zeta$:

$$\hat{Y}_i(\zeta) = 2 \left[\!\!\left[ \left( \sum_{i \in I} [\![ \hat{y}_i > 0 ]\!] \right) \geq \zeta \right]\!\!\right] - 1. \tag{2}$$

This formulation is clearly more robust, as a certain number of false positive instances is tolerated. Standard MIL corresponds in this notation to the case $\zeta = 1$. We have considered and evaluated two options to estimate $\zeta$ during training:

1. **estimation** – estimate the ratio of positive instances in positive and negative bags and set $\zeta$ as their average.
2. **optimization** – choose $\zeta$ minimizing the number of bag-level classification errors.

$$\hat{\zeta} = \arg\min \sum\nolimits_{I \in \mathcal{I}} \mathcal{L}_{\text{agg}}(B_I, \zeta), \tag{3}$$

with the bag-loss function

$$\mathcal{L}_{\text{agg}}(B_I, \zeta) = [\![ Y_I^* \cdot \hat{Y}_I(\zeta) < 0 ]\!] \tag{4}$$

If there are multiple choices for $\zeta_{\text{opt}}$ minimizing the bag-loss, they must all lie within a range $[\min \zeta_{\text{opt}}, \max \hat{\zeta_{\text{opt}}}]$ and we set $\hat{\zeta}$ as their mean to maximize the distance from both negative and positive counts in such case.

Both estimation and optimization are performed in each iteration and can be easily combined with the MIL algorithms mentioned above. Those methods force in each iteration the 'top' instance to be positive for each positive bag. We set $k$ 'top' instances to be positive, with $k = \zeta$. The standard MIL heuristics use the parameter implicitly with $k = 1$. A similar MIL generalization was suggested by Li *et al.*,[11] which is based on averaging top $k$ instances (**top-MI-SVM**) from each bag and then training an SVM using those averages, with $k$ chosen by cross-validation.

Table 1: Listing of hyper-parameters for the different classifier classes used in the experiments. Note that $k$ applies only to top-MI-SVM.

| SVM-based | | RF-based | |
|---|---|---|---|
| $C$ | $\{1, 10.0, 100.0\}$ | $N_{trees}$ | $\{10, 15, 20\}$ |
| $\gamma_{\text{rbf}}$ | $\{0.1, 1.0, 10.0\}$ | max. depth | $\{10, 20\}$ |
| $k$ | $\{1, 5, 10\}$ | | |

Table 2: AUC and $F_1$ scores for Experiment 2. Each column shows the measure (mean±sd) for the standard variant ($k = 1$) and the two proposed extensions ($\zeta_{\text{est}}$, $\zeta_{\text{opt}}$). Significant differences with respect to the standard variant ($k = 1$) are indicated with $^{\dagger}p < 0.01$ and $^{\ddagger}p < 0.001$. Best score is shown in bold.

| | k | mi-SVM[9] | MI-SVM[9] | top-MI-SVM[11] | MIL-RF[10] | MIL-ARF[10] |
|---|---|---|---|---|---|---|
| AUC | 1 | $0.979 \pm 0.019$ | $0.960 \pm 0.025$ | $0.943 \pm 0.040$ | $0.974 \pm 0.018$ | $0.970 \pm 0.027$ |
| | $\zeta_{\text{est}}$ | $0.978 \pm 0.020$ | $0.959 \pm 0.030$ | $0.622 \pm 0.204$ | $\mathbf{0.980 \pm 0.015}^{\dagger}$ | $0.961 \pm 0.031$ |
| | $\zeta_{\text{opt}}$ | $0.978 \pm 0.020$ | $0.963 \pm 0.025$ | $0.634 \pm 0.214$ | $0.978 \pm 0.018$ | $0.975 \pm 0.018$ |
| $F_1$ | 1 | $0.903 \pm 0.019$ | $0.949 \pm 0.024$ | $0.874 \pm 0.106$ | $0.916 \pm 0.022$ | $0.909 \pm 0.024$ |
| | $\zeta_{\text{est}}$ | $\mathbf{0.983 \pm 0.013}^{\ddagger}$ | $0.958 \pm 0.016$ | $0.872 \pm 0.114$ | $0.961 \pm 0.034^{\ddagger}$ | $0.973 \pm 0.016^{\ddagger}$ |
| | $\zeta_{\text{opt}}$ | $\mathbf{0.983 \pm 0.013}^{\ddagger}$ | $0.959 \pm 0.025$ | $0.872 \pm 0.113$ | $0.968 \pm 0.020^{\ddagger}$ | $0.968 \pm 0.019^{\ddagger}$ |

## 3. RESULTS

Our dataset consists of 220 subjects and was divided into three groups: 52 healthy controls (group A), 78 subjects with monoclonal plasma cell disorder (MPCD) but without observed bone marrow infiltration (group B), and 90 subjects with MPCD and infiltrations (group C). The image acquisition was performed on a 256-slice scanner (Brilliance iCT 256; Philips Healthcare, Best, The Netherlands) at a matrix-size of 512 × 512 and a voxel size of 0.976 × 0.976 × 0.450 mm.

The same three binary classification experiments as in the baseline method[4] and which cover different scenarios of assigning the data groups (A, B and C) to positive and negative classes were carried out: (1) $A$ vs. $C$; (2) $A$ vs. $B + C$ ; and (3) $A + B$ vs. C. The experiments differ in handling the group $B$ which may contain people without MM but also people with MM, which has not yet manifested itself clearly in the bones.

We compare the MIL-methods *mi-SVM, MI-SVM, top-MI-SVM, MIL-ARF and MIL-RF* in their standard form and their extended versions using either the estimation or optimization strategy for establishing the decision threshold against the previous, baseline method.[4] We first perform a hyper-parameter grid search (Table 1) to estimate the best parameters for each experiment. We then fix these to the mean values estimated in the first run and evaluate the classifier with stratified random bootstrap sampling and 30-fold repetition. The local intensity features are computed for $N_h = 16$ histogram bins over the range $[-300\text{HU}, 250\text{HU}]$ in each ROI of the grid parametrized with $N_z = 10, N_t = 2, N_\phi = 3$.

We evaluate the ROC curve by varying the threshold $\zeta$ and calculate the AUC criterion. We also calculate the $F_1$ measure, choosing a working point minimizing the total number of false positives and negatives.

For Experiment 2, the generalized versions exhibited an increase of the $F_1$-score, with significant improvements for the instance-based classifier (mi-SVM, MIL-RF and MIL-ARF), while the performance regarding ROC-AUC remained almost the same (see Table 2). Further, as we can see in Table 3, the results of all standard MIL method on the most difficult Experiment 3 are improved by our suggested generalization. The top-MI-SVM[11] (run with $k = 3$) also works well but the best results in terms of both AUC and $F_1$ are achieved by the MIL-ARF[10] random forest method with optimized $\zeta$. In Table 3 we see that the MIL-ARF+$\zeta_{\text{opt}}$ outperforms the baseline methods for all three experiments and all evaluation metrics. Examples of detected regions are shown in Figure 2(B, C).

Our method has almost the same running time as the original one. For MIL-ARF, $k = 1$ took $15.4 \pm 3.8$ s, $\zeta_{\text{est}}$ took $19.6 \pm 4.3$ s, and $\zeta_{\text{opt}}$ took $14.9 \pm 3.2$ s.

Table 3: AUC and $F_1$ scores for Experiment 3. Each column shows the measure (mean±sd) for the standard variant ($k = 1$) and the two proposed extensions ($\zeta_{\text{est}}$, $\zeta_{\text{opt}}$). Significant differences with respect to the standard variant ($k = 1$) are indicated with $^{\dagger}p < 0.01$ and $^{\ddagger}p < 0.001$. Best score is shown in bold.

| | k | mi-SVM[9] | MI-SVM[9] | top-MI-SVM[11] | MIL-RF[10] | MIL-ARF[10] |
|---|---|---|---|---|---|---|
| AUC | 1 | $0.747 \pm 0.065$ | $0.751 \pm 0.053$ | $0.769 \pm 0.108$ | $0.766 \pm 0.083$ | $0.854 \pm 0.055$ |
| | $\zeta_{\text{est}}$ | $0.867 \pm 0.049^{\ddagger}$ | $0.774 \pm 0.068^{\dagger}$ | $0.759 \pm 0.106$ | $0.823 \pm 0.056^{\ddagger}$ | $0.874 \pm 0.048^{\dagger}$ |
| | $\zeta_{\text{opt}}$ | $0.823 \pm 0.066^{\ddagger}$ | $0.764 \pm 0.058$ | $0.661 \pm 0.178$ | $0.807 \pm 0.073^{\dagger}$ | $\mathbf{0.876 \pm 0.048}^{\ddagger}$ |
| $F_1$ | 1 | $0.665 \pm 0.078$ | $0.667 \pm 0.076$ | $0.604 \pm 0.040$ | $0.661 \pm 0.047$ | $0.703 \pm 0.117$ |
| | $\zeta_{\text{est}}$ | $0.589 \pm 0.093$ | $0.688 \pm 0.087$ | $0.587 \pm 0.017$ | $0.703 \pm 0.078$ | $0.729 \pm 0.056^{\dagger}$ |
| | $\zeta_{\text{opt}}$ | $0.725 \pm 0.068^{\ddagger}$ | $0.683 \pm 0.082$ | $0.583 \pm 0.008$ | $0.678 \pm 0.078$ | $\mathbf{0.767 \pm 0.065}^{\ddagger}$ |

Table 4: Comparing the best generalized MIL classifier MIL-ARF+$\zeta_{\text{opt}}$ against the baseline method[4] using AUC, sensitivity, specificity, and $F_1$. Best scores are shown bold, significant differences are marked with $^{\dagger}p < 0.01$ and $^{\ddagger}p < 0.001$.

| | Experiment 1 | | Experiment 2 | | Experiment 3 | |
|---|---|---|---|---|---|---|
| | baseline | MIL-ARF+$\zeta_{\text{opt}}$ | baseline | MIL-ARF+$\zeta_{\text{opt}}$ | baseline | MIL-ARF+$\zeta_{\text{opt}}$ |
| AUC | $0.924 \pm 0.057$ | $\mathbf{0.982 \pm 0.028}^{\ddagger}$ | $0.906 \pm 0.060$ | $\mathbf{0.975 \pm 0.024}^{\ddagger}$ | $0.845 \pm 0.042$ | $\mathbf{0.876 \pm 0.048}^{\ddagger}$ |
| Sens. | $0.981 \pm 0.027$ | $\mathbf{0.991 \pm 0.020}$ | $0.934 \pm 0.040$ | $\mathbf{0.979 \pm 0.018}^{\ddagger}$ | $0.756 \pm 0.117$ | $\mathbf{0.800 \pm 0.143}^{\dagger}$ |
| Spec. | $0.853 \pm 0.101$ | $\mathbf{0.970 \pm 0.053}^{\ddagger}$ | $0.810 \pm 0.158$ | $\mathbf{0.940 \pm 0.067}^{\ddagger}$ | $0.833 \pm 0.095$ | $\mathbf{0.885 \pm 0.062}^{\dagger}$ |
| $F_1$ | $0.952 \pm 0.027$ | $\mathbf{0.965 \pm 0.030}$ | $0.938 \pm 0.023$ | $\mathbf{0.968 \pm 0.019}^{\ddagger}$ | $0.758 \pm 0.055$ | $\mathbf{0.767 \pm 0.065}^{\dagger}$ |

## 4. NEW OR BREAKTHROUGH WORK TO BE PRESENTED

There is only one previous work addressing automatic detection of multiple myeloma from low-dose CT bone images. Our formulation of this task using MIL is novel and outperforms the previous method. The second contribution is a generalization of the MIL formulation and an extension to the aggregation function of the standard MIL solvers. We show that the extended versions of the solver perform better on this task and is likely to be useful for other similar tasks, too.

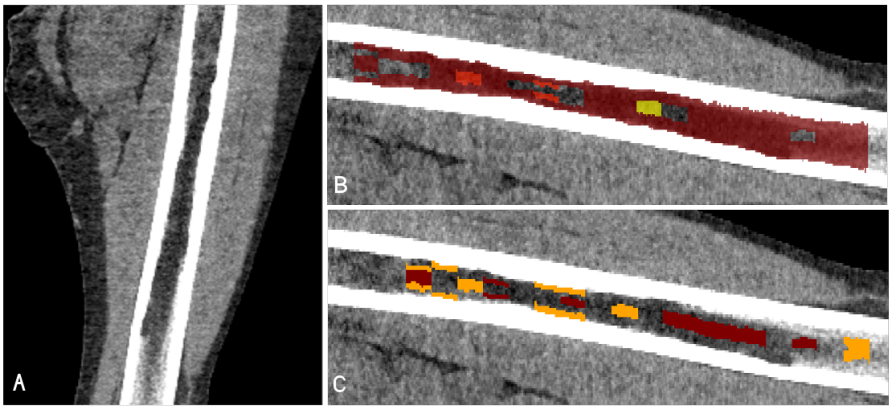


Figure 2: Detected positive ROIs (instances) for the same MM-diseased subject. (A) The subject's femur, (B) predicted positive instances during Experiment 2 and (B) during Experiment 3. All images are displayed with a level-window [-125,250] HU. The ROI color is estimated from the prediction results of 20-fold validation with the MIL-ARF+$\zeta_{\text{opt}}$ classifier – the color is scaled from transparent (predicted label -1 in all cases) to red (predicted label +1 in more than half of the runs).

## 5. CONCLUSION

We have formulated the task of detection of bone marrow infiltrations caused by MM as an MIL task, which enables us to use global labels. We show it to perform better than the previous solution. We have also proposed an extension to an MIL classifier which requires positive bags to contain a certain number of positive instances. This extension improved the results on our problem.

While the results on Experiments 1 and 2 are almost perfect, Experiment 3 remains challenging. This might be caused by the low quality of the training data — for some cases it is difficult even for expert radiologists to reach a consensus whether there is an infiltration or not. In spite of that, we believe that the algorithm could already be useful in the screening setting.

This is an original work which was not published, nor submitted for publication elsewhere.

## 6. ACKNOWLEDGMENTS

This work was supported by the Czech Science Foundation project 17-15361S.

## REFERENCES

[1] Chantry, A., Kazmi, M., Barrington, S., Goh, V., Mulholland, N., Streetly, M., Lai, M., Pratt, G., and British Society for Haematology Guidelines, "Guidelines for the use of imaging in the management of patients with myeloma," *Br. J. Haematol.* (July 2017).

[2] Kyle, R. A. and Rajkumar, S. V., "Criteria for diagnosis, staging, risk stratification and response assessment of multiple myeloma," *Leukemia* **23**, 3–9 (Oct. 2008).

[3] Lambert, L., Ourednicek, P., Meckova, Z., Gavelli, G., Straub, J., and Spicka, I., "Wholebody lowdose computed tomography in multiple myeloma staging: Superior diagnostic performance in the detection of bone lesions, vertebral compression fractures, rib fractures and extraskeletal findings compared to radiography with similar radiation exposure," *Oncology Letters* **13**, 2490–2494 (Apr. 2017).

[4] Martínez-Martínez, F., Kybic, J., Lambert, L., and Mecková, Z., "Fully Automated Classification of Bone Marrow Infiltration in Low-Dose CT of Patients with Multiple Myeloma Based on Probabilistic Density Model and Supervised Learning," *Computers in Biology and Medicine* **71**, 57–66 (Apr. 2016).

[5] Wang, S. and Summers, R., "Machine learning and radiology," *Medical Image Analysis* **16**(5), 933–951 (2012).

[6] Quellec, G., Lamard, M., Abràmoff, M. D., Decencière, E., Lay, B., Erginay, A., Cochener, B., and Cazuguel, G., "A Multiple-Instance Learning Framework for Diabetic Retinopathy Screening," *Medical Image Analysis* **16**, 1228–1240 (Aug. 2012).

[7] Quellec, G., Lamard, M., Cozic, M., Coatrieux, G., and Cazuguel, G., "Multiple-Instance Learning for Anomaly Detection in Digital Mammography," *IEEE Transactions on Medical Imaging* **35**, 1604–1614 (July 2016).

[8] Melendez, J., van Ginneken, B., Maduskar, P., Philipsen, R. H. H. M., Reither, K., Breuninger, M., Adetifa, I. M. O., Maane, R., Ayles, H., and Sánchez, C. I., "A Novel Multiple-Instance Learning-Based Approach to Computer-Aided Detection of Tuberculosis on Chest X-Rays," *IEEE Transactions on Medical Imaging* **34**, 179–192 (Jan. 2015).

[9] Andrews, S., Tsochantaridis, I., and Hofmann, T., "Support Vector Machines for Multiple-Instance Learning," in [*Advances in Neural Information Processing Systems*], 561–568 (2002).

[10] Leistner, C., Saffari, A., and Bischof, H., "MIForests: Multiple-Instance Learning with Randomized Trees," in [*Computer Vision - Eccv 2010, Pt Vi*], Daniilidis, K., Maragos, P., and Paragios, N., eds., **6316**, 29–42, Springer-Verlag Berlin, Berlin (2010).

[11] Li, W. and Vasconcelos, N., "Multiple instance learning for soft bags via top instances," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 4277–4285 (2015).