

Unsupervised Task Design to Meta-Train Medical Image Classifiers ^{*}

Gabriel Maicas[†] Cuong Nguyen[†] Farbod Motlagh[†]
Jacinto C. Nascimento^{‡‡} Gustavo Carneiro[†]

[†]Australian Institute for Machine Learning, The University of Adelaide
^{‡‡}Institute for Systems and Robotics, Instituto Superior Tecnico, Portugal

Abstract. Meta-training has been empirically demonstrated to be the most effective pre-training method for few-shot learning of medical image classifiers (i.e., classifiers modeled with small training sets). However, the effectiveness of meta-training relies on the availability of a reasonable number of hand-designed classification tasks, which are costly to obtain, and consequently rarely available. In this paper, we propose a new method to unsupervisedly design a large number of classification tasks to meta-train medical image classifiers. We evaluate our method on a breast dynamically contrast enhanced magnetic resonance imaging (DCE-MRI) data set that has been used to benchmark few-shot training methods of medical image classifiers. Our results show that the proposed unsupervised task design to meta-train medical image classifiers builds a pre-trained model that, after fine-tuning, produces better classification results than other unsupervised and supervised pre-training methods, and competitive results with respect to meta-training that relies on hand-designed classification tasks.

Keywords: meta-training, unsupervised learning, unsupervised task design, breast image analysis, magnetic resonance imaging, few-shot, pre-training, clustering.

1 Introduction

The accuracy and robustness of deep learning based medical image classifiers is generally positively correlated with the size of the annotated training set used during the modelling process [1]. However, large annotated training sets are expensive and not readily available for some medical image analysis applications, such as breast screening from DCE-MRI [2]. Therefore, training medical image classifiers with small annotated training sets has become a highly investigated topic, particularly after the advent of deep learning [1].

The most competitive medical image classifiers are currently based on convolutional neural networks (CNNs) [1] that need large training sets to be properly modelled. To reduce the need for such large annotated sets, pre-training approaches have been explored in medical image analysis, where the most relevant for our paper are: 1) supervised pre-training using independent data sets [5], where the model is pre-trained by solving a classification problem in a different data set; 2) unsupervised pre-training using clustering [3], where the model is

^{*} Supported by Australian Research Council through grant DP180103232.

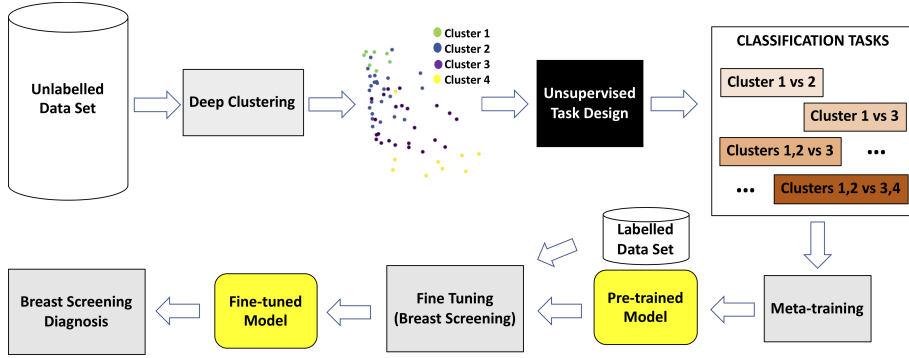


Fig. 1: Unsupervised task design to meta-train medical image classifiers. Deep clustering [3] produces a set of clusters that are used in the unsupervised design of classification tasks. These tasks are used in a meta-training process to produce a pre-trained model that can be fine-tuned to new classification tasks using small labelled training sets, in this paper represented by the breast screening problem from DCE-MRI [4].

pre-trained by performing clustering without any knowledge about the ground truth labels; and 3) unsupervised pre-training using input reconstruction [6], where the model is pre-trained by reconstructing the input images of the training set. Arguably, the main issue with these pre-training methods is that their objective functions are irrelevant for the medical image classifier being developed downstream. Alternatively, the need for pre-training methods can be alleviated with the use of other types of training methods, such as multiple instance learning (MIL) [7] or multi-task learning [8], but both methods still need large training sets. More recently, a pre-trained model produced by supervised meta-training (i.e., a meta-training process that depends on hand-designed classification tasks) showed superior performance compared to the previously described pre-training methods [4]. Nevertheless, these promising meta-training results are counterbalanced by an unappealing need of an expensive hand-designing process to produce the classification tasks [4]. Given the high cost of this process, the availability of a large number of hand-designed classification tasks is rare, which hampers the exploration of meta-training for medical image classifiers.

In this paper, we propose a new method to unsupervisedly produce a large number classification tasks to meta-train medical image classifiers. To this end, we use deep clustering [3] to automatically build image clusters that can be grouped in different ways to enable the design of multiple classification tasks employed in the meta-training process – see Fig. 1. We evaluate our method on the breast screening classification task from a breast DCE-MRI data set that has been used to benchmark few-shot training algorithms of medical image classifiers [4]. Results show that our proposed approach produces classification results that are significantly better than other unsupervised and supervised pre-training methods, and competitive to supervised meta-training.

2 Literature Review

DCE-MRI is a recommended image modality in breast screening programs for patients at high-risk [9]. However, DCE-MRI interpretation is time-consuming and prone to high inter-observer variability [10]. Thus, computer-aided diagnosis (CAD) systems are being developed to assist radiologists increase their diagnosis sensitivity [11] and specificity [12], and reduce analysis time. However, the development of CAD systems for breast DCE-MRI is challenging due in part to the small size of annotated data sets available for training.

Meta-training has been shown to be an effective strategy to improve the learning of classifiers using relatively small training sets [13]. For instance, Maicas *et al.* [4] proposed the use of hand-designed breast classification tasks to meta-train a model that was then fine-tuned to solve the breast screening task. Results showed that this method improves over other strategies to train classifiers from small data sets, such as MIL [7] and multi-task learning [8]. However, the method proposed in [4] relies on costly hand-designed classification tasks.

Similarly to our paper, Hsu *et al.* [14] proposed an unsupervised method to design computer vision classification tasks for meta-training. Results showed that this approach produced worse classification performance than meta-training modelled with hand-designed tasks (i.e., supervised meta-training). We believe that the reason behind this drop in performance lies in the large number of hand-designed tasks already available for supervised meta-training in computer vision applications [14], enabling a good classification performance baseline. The difficulty to obtain a large number of hand-designed tasks for medical image classification problems means that the number of these hand-designed tasks will be small, which may result in a relatively low classification performance baseline. We hypothesize that our proposed method that unsupervisedly designs a large number of classification tasks to meta-train a medical image classifier can achieve a classification performance that is at least comparable to supervised meta-training [4] trained with a small number of hand-designed tasks. Our proposed method has the advantage that it does not rely on costly hand-designed tasks.

3 Data Set and Methods

3.1 Data set

The data set is represented by $\mathcal{D} = \{(\mathbf{v}_i, \mathbf{t}_i, b_i, y_i)\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{v} : \Omega \rightarrow \mathbb{R}$ corresponds to the first DCE-MRI subtraction volume (Ω denotes the volume lattice) [15], $\mathbf{t} : \Omega \rightarrow \mathbb{R}$ represents the T1-weighted MRI only used to separate the left and breast regions of the volume, $b \in \{\text{left}, \text{right}\}$ indicates the left or right breast, and $y \in \mathcal{Y} = \{0, 1\}$ indicates the classification label: no malignant findings, or malignant findings, respectively.

3.2 Deep Clustering to Unsupervisedly Design Classification Tasks

The proposed unsupervised task design method builds several binary classification tasks from image groups formed by deep clustering [3]. The training of deep clustering alternates an optimisation of two objective functions [3]. We denote the θ -parameterised model that produces the unsupervised learning features by

$f_\theta(\mathbf{v}) \in \mathbb{R}^D$ and the ω -parameterised classifier that produces a pseudo-label representing one of the unknown K classes and is placed on top of $f_\theta(\cdot)$ by $g_\omega(f_\theta(\mathbf{v})) \in \{0, 1\}^K$. The first objective function is the cross-entropy loss $\ell(\cdot)$ with respect to the pseudo-labels $\{\tilde{\mathbf{y}}_i\}_{i=1}^{|\mathcal{D}|}$, with $\tilde{\mathbf{y}} \in \tilde{\mathcal{Y}} = \{0, 1\}^K$,

$$\min_{\theta, \omega} \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \ell(g_\omega(f_\theta(\mathbf{v}_i)), \tilde{\mathbf{y}}_i), \quad (1)$$

which is used to estimate the optimal θ^* and ω^* . The second objective function finds the K centroids, denoted by $\mathbf{C} \in \mathbb{R}^{D \times K}$, and pseudo-labels $\tilde{\mathbf{y}}$ with

$$\min_{\mathbf{C}} \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \min_{\tilde{\mathbf{y}}_i} \|f_\theta(\mathbf{v}_i) - \mathbf{C}\tilde{\mathbf{y}}_i\|_2^2, \quad (2)$$

where $\tilde{\mathbf{y}}_i$ is a K -dim one-hot vector.

Each step of the optimization above will generate new values for the model parameters, centroids and pseudo-labels. We extend deep clustering [3] with a model selection process based on maximising the Silhouette coefficient that measures clustering quality [16] with

$$\kappa = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \frac{b(i) - a(i)}{\max(a(i), b(i))}, \quad (3)$$

where $a(i)$ represents the average ℓ_2 distance between $f_\theta(\mathbf{v}_i)$ and all points $f_\theta(\mathbf{v}_j)$ where $i \neq j$ and $\tilde{\mathbf{y}}_i = \tilde{\mathbf{y}}_j$; and $b(i)$ denotes the smallest average ℓ_2 distance between $f_\theta(\mathbf{v}_i)$ and $f_\theta(\mathbf{v}_j)$ where $i \neq j$ and $\tilde{\mathbf{y}}_i \neq \tilde{\mathbf{y}}_j$.

The unsupervised design of classification tasks is based on the formation of L binary classification problems derived from the pseudo-labels obtained from (2). Each of these L binary classification problems is built by randomly selecting 2 nonempty and disjoint subsets $\mathcal{K}_l^{(0)}$ and $\mathcal{K}_l^{(1)}$ from the pseudo label set $\{1, 2, \dots, K\}$ and labelling their corresponding data points as class 0 and 1, respectively. Note that the number of classification tasks for a given K is $L = \sum_{i=1}^{n-1} \sum_{k=1}^{\min(i, n-i)} \frac{\binom{n}{i} \times \binom{n-i}{k}}{1 + \delta(i-k)}$, where $\binom{A}{B}$ denotes the binomial coefficient, and $\delta(\cdot)$ represents the Dirac delta function.

3.3 Meta-training with the Unsupervised Classification Tasks

Meta-training estimates the parameters of a meta-learner, so it can be used as a pre-trained model that is efficiently fine-tuned to previously unseen classification tasks, using small annotated training sets [13]. The algorithm assumes that there exists a task distribution \mathcal{T} , from which each classification task \mathcal{T}_l is drawn, where each task comprises a training set $\{\mathbf{v}_i^{(l,t)}, \tilde{\mathbf{y}}_i^{(l,t)}\}_{i=1}^M$ and a testing set $\{\mathbf{v}_i^{(l,v)}, \tilde{\mathbf{y}}_i^{(l,v)}\}_{i=1}^N$, with $M \ll N$ and $M + N = |\mathcal{T}_l|$. Meta-training iteratively samples T tasks from \mathcal{T} , and re-trains a multi-target classifier for those tasks using the training and testing sets defined above.

We use the MAML meta-training [17] that consists of a Bayesian hierarchical model, where ψ denotes the classifier meta parameter, and ϕ_l represents the

parameter for task \mathcal{T}_l . The meta-training objective function is defined by:

$$\max_{\psi} \log p(\mathcal{Y}_{l=1..T}^{(v)} | \mathcal{Y}_{l=1..T}^{(t)}, \mathcal{V}_{l=1..T}^{(v)}, \mathcal{V}_{l=1..T}^{(t)}, \psi), \quad (4)$$

where T is the number of tasks per meta-training iteration, $\mathcal{Y}_l^{(v)} = \{\tilde{y}_i^{(l,v)}\}_{i=1}^N$, $\mathcal{Y}_l^{(t)} = \{\tilde{y}_i^{(l,t)}\}_{i=1}^M$, $\mathcal{V}_l^{(v)} = \{\mathbf{v}_i^{(l,v)}\}_{i=1}^N$, and $\mathcal{V}_l^{(t)} = \{\mathbf{v}_i^{(l,t)}\}_{i=1}^M$. In (4), we have

$$\log p(\mathcal{Y}_{l=1..T}^{(v)} | \mathcal{Y}_{l=1..T}^{(t)}, \mathcal{V}_{l=1..T}^{(v)}, \mathcal{V}_{l=1..T}^{(t)}, \psi) \geq \sum_{l=1}^T \mathbb{E}_{p(\phi_l | \mathcal{Y}_l^{(t)}, \mathcal{V}_l^{(t)}, \psi)} \left[\log p(\mathcal{Y}_l^{(v)} | \mathcal{V}_l^{(v)}, \phi_l) \right], \quad (5)$$

where the lower bound is derived from Jensen’s inequality [18]. Therefore, the maximisation in (4) is approximated with the lower bound maximisation in (5), where the posterior $p(\phi_l | \mathcal{Y}_l^{(t)}, \mathcal{V}_l^{(t)}, \psi)$ is approximated with a Dirac delta function at a local optimal task-specific model parameter ϕ_l^* , with $p(\phi_l | \mathcal{Y}_l^{(t)}, \mathcal{V}_l^{(t)}, \psi) = \delta(\phi_l - \phi_l^*)$. The local optimal model parameter ϕ_l^* is obtained with truncated gradient descent initialised by the meta parameters ψ :

$$\phi_l^* = \psi - \alpha \nabla_{\phi_l} \left[-\log p(\mathcal{Y}_l^{(t)} | \mathcal{V}_l^{(t)}, \phi_l) \right], \quad (6)$$

where α is the learning rate, and the truncated gradient descent consists of a single step of (6). Maximising the lower bound of the log likelihood in (5) represents the MAML algorithm in [13], which produces a pre-trained model that can quickly learn new tasks drawn from \mathcal{T} .

4 Experiments and Results

4.1 Experimental Set-Up

We evaluate our proposed method on a breast DCE-MRI data set [2] (formally defined in Sec. 3.1), which has previously been used to evaluate few-shot training methods [4]. To allow a fair comparison with previous papers, we split the data set in a patient-wise manner into the same training, validation and testing sets, containing 45, 13, and 59 patients, respectively. We use the T1-weighted MRI to automatically extract the left and right breast regions from the first DCE-MRI subtraction volume [4]. Each breast region is resized into a volume of $100 \times 100 \times 50$ [4]. For the breast screening problem, only breasts that contain a malignant finding(s) are considered positive, while breasts with only benign findings or no findings are considered negative. There are 30, 9, and 38 positive and 60, 17, and 80 negative breasts in the training, validation and testing sets, respectively.

The model $f_{\theta}(\mathbf{v})$ that unsupervisedly produces the volume features is a 3D Densenet [19] composed of five dense blocks, each containing two dense layers. The features represent the input to the deep clustering algorithm, explained in Sec.3.2, with the number of clusters $K \in \{3, 4, 5\}$. The model that is meta-trained, and fine-tuned, has the same architecture as $f_{\theta}(\cdot)$. During meta-training, we use a meta learning rate $\alpha = 0.001$ in (6). At each meta-iteration, a meta-batch size of $T = 4$ classification tasks is sampled according to a random or a curriculum learning strategy [4]. The meta-trained model is fine-tuned to the

breast screening task using the entire training set, where model selection is performed using the validation set and results are reported in the test set.

The evaluation for the breast screening problem is based on the area under the ROC curve (AUC). We also measure the standard error utilising an estimate based on the Wilcoxon test [20] that estimates confidence intervals based on the testing set. In this evaluation, we study the type of task sampling for meta-training, i.e. random, or curriculum learning [4], and the influence of the number of clusters K in (1), used to build the tasks. We compare our method (U-MT) with the previously proposed supervised meta-training for the case where the breast screening task is included (S-MT (S)) and not included (S-MT (NS)) in the meta-training process. We also compare our method with: a) Densenet trained from scratch on the breast screening task; b) Densenet from (a) fine-tuned with MIL [7]; c) Densenet trained with multi-tasking (using hand-designed tasks) [4]; d) Densenet pre-trained as a variational autoencoder (i.e., unsupervised training) and fine-tuned for the breast screening task; and e) Densenet pre-trained with deep clustering (i.e., unsupervised training) and fine-tuned for the breast screening task. All Densenet models of these competing methods have the same architecture as the meta-trained model described above. The rationale for baselines (d) and (e) is to evaluate the effect of pre-training based on a reconstruction or a clustering scheme. With this purpose, we present results based on nearest neighbor classification and the fine-tuned classification model.

4.2 Results

We show the AUC results (\pm standard error) for breast screening baselines in Tab. 1. Table 2 presents the results of meta-training, as a function of $K \in \{3, 4, 5\}$, with supervised and unsupervised task design using random and curriculum learning task sampling methods. Figure 2 presents examples of breast screening classification.

Training Method Baseline	AUC
From Scratch [19]	0.83 ± 0.04
MIL based fine-tuning [7]	0.85 ± 0.04
Multi-Task [8]	0.85 ± 0.04
Variational Autoencoder + Nearest Neighbour	0.61 ± 0.06
Variational Autoencoder + Fine-Tune in breast screening	0.84 ± 0.04
Deep Clustering + Nearest Neighbour	0.53 ± 0.06
Deep Clustering + Fine-Tune in breast screening	0.80 ± 0.05

Table 1: AUC results (\pm standard error) for breast screening baselines.

We measure the statistical significance of the difference in performance between our best performing approaches (Random with $K = 5$ and Curriculum with $K = 5$) and all baseline methods, obtaining a p-value $p \leq 0.001$ for all cases (unpaired two-tailed t-test). Also, comparing our newly proposed U-MT (Random with $K = 5$) and S-MT (S) (Curriculum with $K = 3$) [4], we obtain a p-value $p > 0.05$.

	Random			Curriculum		
	$K = 3$	$K = 4$	$K = 5$	$K = 3$	$K = 4$	$K = 5$
S-MT [4] (S)	0.86 ± 0.04	N/A	N/A	0.90 ± 0.04	N/A	N/A
S-MT [4] (NS)	0.85 ± 0.04	N/A	N/A	0.89 ± 0.04	N/A	N/A
U-MT (Ours)	0.81 ± 0.05	0.88 ± 0.04	0.89 ± 0.04	0.87 ± 0.04	0.86 ± 0.04	0.88 ± 0.04

Table 2: AUC for the breast screening task for our proposed method (U-MT) as a function of the number of image clusters K and the task sampling method (random and curriculum). We also present the results of supervised meta-training [4] (S-MT) for the cases where the breast screening is included (labelled as S) and not included (labelled as NS) in the meta-training tasks. N/A indicates that the experiment is not feasible due to the lack of extra ground truth labels.

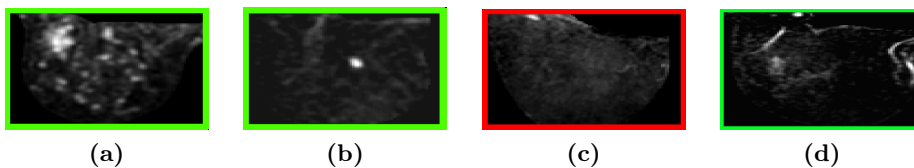


Fig. 2: Example of breast screening diagnosis produced by our approach. Image (2a) shows the correct positive diagnosis of a breast containing a malignant tumour. Image (2b) shows the correct negative diagnosis of a breast with a benign tumour. Image (2c) shows the incorrect positive classification of a breast containing no tumours. Image (2d) shows the correct negative diagnosis of a breast with a benign tumour.

5 Discussion and Conclusion

We have presented a new method that unsupervisedly designs classification tasks to meta-train medical image classifiers. Our method significantly outperforms several baselines consisting of traditional pre-training methods based on variational autoencoder, deep clustering, MIL, and multi-task learning (see Tab. 1). Our method also produces results comparable to the state-of-the-art set by meta-training using hand-designed tasks [4] (see Tab. 2). However, instead of using manually defined labels during meta-training, we unsupervisedly build classification tasks, allowing us to build a larger set of tasks, compared to the hand-designed ones. Also from Tab. 2, we notice that larger number of tasks, which increases with the number of clusters (Sec. 3.2), generally implies better AUC results. This confirms our initial hypothesis that, differently from computer vision problems, automatically building tasks is of great importance for medical image classification problems, where image labels that allow a large number of tasks are costly to obtain. We also observe that sampling tasks according to curriculum learning provides a good improvement of accuracy compared to random task sampling for a small number of clusters ($K = 3$), but not for larger number of tasks ($K = 5$). We hypothesize that meta-training with curriculum learning sampling needs a larger number of meta-iterations to learn a curriculum that is better than random task sampling. Given the large number of tasks for

$K \in \{4, 5\}$, the meta-training process converged before the curriculum learning algorithm – that deserves further research.

References

1. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. *Medical image analysis* (2017)
2. McClymont, D., Mehnert, A., Trakic, A., Kennedy, D., Crozier, S.: Fully automatic lesion segmentation in breast mri using mean-shift and graph-cuts on a region adjacency graph. *JMRI* (2014)
3. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: *ECCV*. (2018)
4. Maicas, G., Bradley, A.P., Nascimento, J.C., Reid, I., Carneiro, G.: Training medical image analysis systems like radiologists. In: *MICCAI*. (2018)
5. Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E., Greenspan, H.: Chest pathology detection using deep learning with non-medical training. In: *ISBI*. (2015)
6. Dong, L.F., Gan, Y.Z., Mao, X.L., Yang, Y.B., Shen, C.: Learning deep representations using convolutional auto-encoders with symmetric skip connections. In: *ICASSP*. (2018)
7. Zhu, W., Lou, Q., Vang, Y.S., Xie, X.: Deep multi-instance networks with sparse label assignment for whole mammogram classification. In: *MICCAI*. (2017)
8. Xue, W., Brahm, G., et al.: Full left ventricle quantification via deep multitask relationships learning. *Medical image analysis* (2018)
9. Mainiero, M.B., Moy, L., Baron, P., Didwania, A.D., Green, E.D., Heller, S.L., Holbrook, A.I., Lee, S.J., Lewin, A.A., Lourenco, A.P., et al.: Acr appropriateness criteria® breast cancer screening. *JACR* (2017)
10. Grimm, L.J., Anderson, A.L., Baker, J.A., Johnson, K.S., Walsh, R., Yoon, S.C., Ghate, S.V.: Interobserver variability between breast imagers using the fifth edition of the bi-rads mri lexicon. *American Journal of Roentgenology* (2015)
11. Vreemann, S., Gubern-Merida, A., Lardenoije, S., Bult, P., Karssemeijer, N., Pinker, K., Mann, R.M.: The frequency of missed breast cancers in women participating in a high-risk mri screening program. *Breast Cancer Research and Treatment* (2018)
12. Meinel, L.A., Stolpen, A.H., Berbaum, K.S., Fajardo, L.L., Reinhardt, J.M.: Breast mri lesion classification: Improved performance of human readers with a backpropagation neural network computer-aided diagnosis (cad) system. *JMRI* (2007)
13. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: *ICML*. (2017)
14. Hsu, K., Levine, S., Finn, C.: Unsupervised learning via meta-learning. In: *ICLR*. (2019)
15. Gilbert, F., Selamoglu, A.: Personalised screening: is this the way forward? *Clinical radiology* (2018)
16. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* (1987)
17. Grant, E., Finn, C., Levine, S., Darrell, T., Griffiths, T.: Recasting gradient-based meta-learning as hierarchical bayes. (2018)
18. Bishop, C.M.: *Pattern recognition and machine learning*. springer (2006)
19. Huang, G., Liu, Z.: Densely connected convolutional networks. In: *CVPR*. (2017)
20. Bradley, A.P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition* (1997)