Noname manuscript No.
(will be inserted by the editor)

DTCo: An ensemble SSL algorithm for X-rays classification

Ioannis Livieris · Theodore Kotsilieris · Ioannis Anagnostopoulos · Vassilis Tampakas

the date of receipt and acceptance should be inserted later

Abstract During the last decades, the classification of images constitutes a typical method for diagnosing many abnormalities and diseases. The application of an efficient classification method is considered essential in modern diagnostic medicine in order to increase the number of analyzed patients and decrease the analysis time. The significant storage capabilities of electronic media have enabled research centers to accumulate repositories of classified (labeled) images and mostly of a large number of unclassified (unlabeled) images. Semi-supervised learning algorithms have become a hot topic of research as an alternative to traditional classification methods, exploiting the explicit classification information of labeled data with the knowledge hidden in the unlabeled data for building powerful and effective classifiers. In this work, we propose a new ensemble self-labeled algorithm, called DTCo, for X-rays classification. The efficacy of the presented algorithm is illustrated by a series of experiments against other state-of-the-art self-labeled methods.

Keywords Semi-supervised learning · self-labeled algorithms · ensemble learning · X-ray classification · lung abnormalities.

1 Introduction

Nowadays, machine learning and data mining have emerged as widely accepted techniques in the field of diagnostic medicine. They have received much attention in solving medical diagnostic tasks as they encompass the following attributes: good performance, ability to deal with missing and noisy data,

I.E. Livieris
Department of Computer & Informatics Engineering (DISK Lab), Technological Educational Institute of Western Greece, Greece, GR 263-34. E-mail: livieris@teiwest.gr

T. Kotsilieris
Department of Business Administration (LAIQDA Lab), Technological Educational Institute of Peloponnese, Greece, GR 241-00. E-mail: tkots@teikal.gr

I. Anagnostopoulos
Department of Computer Science and Biomedical Informatics, University of Thessaly, Greece, GR 351-00. E-mail: janag@dib.uth.gr

V. Tampakas
Department of Computer & Informatics Engineering (DISK Lab), Technological Educational Institute of Western Greece, Greece, GR 263-34. E-mail: vtampakas@teimes.gr

transparency of diagnostic knowledge and explanatory decisions (Kononenko 2001). Furthermore, commonly employed data analytic tools pose several limitations to the complex environment of health data analysis.

Research efforts have been devoted on the development of intelligent computational systems that efficiently analyze medical data and images in order to extract useful knowledge in the field of pulmonary diseases and disorders. It worths mentioning that, as reported by the World Health Organization (2017), tuberculosis (TB) is one of the top 10 causes of death, caused approximately 1.6 million deaths only in 2017 while millions of people fall sick with TB each year. At the same time, pneumonia accounts for 16% of all children aged 0-5 years while lung cancer is the sixth more common cause of death worldwide claiming 1.7 million lives in 2016. Although several diagnostic tests are widely employed in the pulmonary diseases domain, their application on a large scale is usually cumbersome, costly, time consuming to process and prone to human errors (i.e. diagnosis errors). Therefore, lung diagnosis domain has been benefited extensively by the advances of machine learning and data mining techniques over posterior-anterior Chest X-Rays (CXRs) in order to analyze the suspected region and search for any abnormalities. CXR imaging is being widely applied for diagnosis due to low cost and easy operation. Although the interpretation of such medical images is usually performed by experts (i.e. radiologists and lung specialists), recent advances in medical informatics has shifted the interest to the development of computer based decision support and diagnosis systems.

The rapid advances in digital chest radiography and the significant storage capabilities of electronic media, have enabled research centers to accumulate large repositories of classified (labeled) images and mostly of unclassified (unlabeled) images from human experts. Therefore, researchers and medical stuff have a significant potential to transform biomedical research and the delivery of healthcare by leveraging and exploit these images. Generally, the progress in the field has been hampered by the lack of available labeled images for efficiently training an accurate supervised classifier. Nevertheless, the process of correctly labeling new unlabeled CXRs frequently requires the efforts of specialized personnel and expert physicians, which will incur high time and monetary costs.

Semi-Supervised Learning (SSL) algorithms constitute the appropriate and effective machine learning methodology for extracting useful knowledge from both labeled and unlabeled data. In contrast to traditional classification approaches, semi-supervised algorithms leverage the large amount of unlabeled data in order to reduce data sparsity in the labeled training data and boost the classifier performance, particularly focusing on the setting where the amount of available labeled data is limited. Thus, these algorithms have received considerable attention due to their potential for reducing the effort of labeling data while still preserving competitive and sometimes better classification performance (see Zhou (2011), Zhu (2011), Zhu & Goldberg (2009) and the references therein).

Self-labeled algorithms are probably considered the most popular class of SSL algorithms, exploiting the unlabeled data via a self-learning process based on supervised prediction models. They perform an iterative procedure, aiming to obtain an enlarged labeled dataset, in which they accept that their own predictions tend to be correct. Recently, Triguero et al. (2015) proposed an in-depth taxonomy based on the main characteristics and conducted an extended study of their classification efficacy on several datasets.

In this work, we propose a new ensemble self-labeled algorithm, called DTCo, for the classification of X-rays. The proposed algorithm combines the predictions of three of the most efficient and frequently used self-labeled algorithms, utilizing a maximum-probability voting scheme. Our preliminary numerical experiments demonstrate the efficacy of DTCo for the detection of abnormalities from X-rays, illustrating that reliable and robust classification models could be developed by the adaptation of ensemble methodologies in the semi-supervised framework.

The remainder of this paper is organized as follows: Section 2 presents a survey of recent studies concerning the application of data mining in X-rays classification. Section 3 presents the proposed ensemble semi-supervised classification algorithm. Section 4 presents a series of experiments in order

to examine and evaluate the accuracy of the proposed algorithm against the most popular self-labeled classification algorithms and Section 5 presents our concluding remarks and future work.

2 Related work

During the last decades, we have witnessed the significance of medical imaging for the diagnosis, the early detection and the treatment of diseases. The advances of digital technology and chest radiography as well as the rapid development of digital image retrieval have renewed the interest and the progress in new technologies for the diagnosis of abnormalities. In particular, there has been a growing interest in developing Computer-Aided Diagnostic (CAD) systems for the detection of abnormalities, therefore a variety of methods has been used aiming on classifying and/or detecting anomalies in medical images. Most of the CAD methods proved to be powerful tools which could assist medical staff in hospitals and lead to better results in diagnosing a patient. However, despite all this effort, there is still no widely utilized method for classifying medical images since the medical domain requires high accuracy; especially the rate of false negatives is imperative very low. Along this line, van Ginneken et al. (2009) presented a survey in which they stated that forty-five years after the initial work on CAD in chest radiology, there are still no systems that can accurately read chest radiographs. To this end, in recent years, a number of rewarding studies has being carried out, some of which are briefly described in the next paragraphs.

Hogeweg et al. (2010) combined a texture-based abnormality detection system with a clavicle detection stage in order to suppress false positive responses. Based on their previous work, Hogeweg et al. (2012) utilized a combination of pixel classifiers and activated shape models for clavicle segmentation. Notice that the clavicle region consists of a notoriously difficult region for the detection of TB since the clavicles can obscure manifestations of TB in the apex of the lung. Bearing in mind, Xu et al. (2011) introduced a novel technique by hybridizing a model-based template matching technique with image enhancement based on the Hessian matrix.

Muyoyeta et al. (2014) illustrated the ability of CAD systems for discriminating CXR as normal or abnormal and the potential for roll-out of digital X-ray technology, especially in high burden settings where human resources are scarce. They utilized data from 350 patients and concluded that CAD and CXR can be used as a pre-screening tool before applying more expensive diagnostic tests. On the other hand, cost effectiveness of such strategies would have to be ascertained. Another similar work is presented by Jaeger et al. (2014) which proposed an approach for detecting tuberculosis in conventional posteroanterior chest radiographs. Initially, their proposed method extracts the lung region from the CXRs utilizing a graph cut segmentation method and a set of texture and shape features in the lung region is computed in order to classify the patient as normal or abnormal. The results of the numerical experiments on two real-world datasets, revealed that the proposed CAD system for TB screening achieved high performance, relevant to that of human readings.

Candemir et al. (2014) presented a non-rigid registration-driven robust lung segmentation method using image retrieval-based patient specific adaptive lung models which detect lung boundaries. More specifically, their suggested methodology incorporates non-grid registration with CXR databases of pre-segmented lung regions to develop an anatomical atlas as a guide combined with graph cuts based on image region refinement. Moreover, their proposed method was evaluated utilizing three different datasets containing in total 585 chest radiographs from patients with normal lungs and various pulmonary diseases indicating the robustness and effectiveness of the proposed approach.

Plankis et al. (2017) developed a CAD system which calculates the lung regions of interest, performs lung image segmentation and automated disease recognition. More specifically, the main tasks of the proposed system comprise: a) recognition and control of radiograph images for later lung

textures analysis, b) Daubechies wavelet transformation, c) 12 texture parameters computation based on wavelet coefficients, d) supervised machine learning for clinical decisions, e) decision evaluation.

Santosh & Antani (2018) developed a novel concept of using right and left lung region changes into account and represent in terms of symmetry, and have automated the chest X-ray system for the evidence of tuberculosis. Their method utilizes common pulmonary abnormalities exhibited in CXR images including cavitations, consolidations, infiltrates, blunted costophrenic angles, opacities, pleural effusion Moreover, to compute lung region symmetry, we have used multi-scale shape features and edge-based and texture-based features (take internal content). Unlike other the state-of-the art techniques, they have proved that the way the features are represented is the appropriate for chest X-ray screening to detect pulmonary abnormalities. They presented some encouraging performance by using voting-based combination of three different classifiers: random forest, artificial neural network and Bayesian network on two CXR benchmarks.

Alam et al. (2018) developed an efficient lung cancer detection and prediction algorithm using multi-class support vector machine classifier. In every stage of classification, the image enhancement and the image segmentation have been done separately. Furthermore, image scaling, color space transformation and contrast enhancement have been utilized for image enhancement while threshold and marker-controlled watershed based segmentation has been utilized for segmentation. Next, a set of textural features extracted from the separated regions of interest is been categorized by the support vector machine classifier. The proposed algorithm can efficiently detect cancer affected cell and the corresponding stage such as initial, middle, or final stage while if no cancer affected cell is found in the input image then it checks the probability of lung cancer.

In more recent works, Livieris et al. (2018) proposed CST-Voting, an ensemble semi-supervised learning algorithm for CXR classification of tuberculosis. Their proposed algorithm exploits the individual predictions of three of the most efficient and frequently used self-labeled algorithms i.e., Co-training, Self-training and Tri-training, using a voting methodology. The authors presented some numerical experiments demonstrating the efficiency of the proposed algorithm against several semi-supervised learning algorithms, and illustrating that reliable and robust prediction models could be developed utilizing a few labeled and many unlabeled data.

3 DTCo algorithm

In this section, we present a detailed description of the proposed self-labeled algorithm for X-ray classification, which is based on a maximum-probability voting scheme.

Motivated by Livieris et al. (2018) and Livieris (2019), we consider to develop an ensemble algorithm based on the idea of generating classifiers by applying different self-labeled algorithms (with heterogeneous model representations) to a single dataset. On this basis, the learning algorithms, which constitute the proposed ensemble, are: Democratic Co-learning, Tri-training and Co-Bagging. The motivation for this selection is based upon the fact that these algorithms have been presented as the most efficient and robust self-labeled algorithms Triguero et al. (2015).

These algorithms are self-labeled ones, which exploit the hidden information in unlabeled data using different methodologies. More specifically, *Democratic Co-learning* algorithm (Zhou & Goldman 2004) follows the multi-view theory but from another aspect based on the idea of ensemble learning and majority voting. More analytically, this algorithm utilizes multiple algorithms for producing the necessary information and endorses a voted majority process for the final decision, instead of asking for more than one views of the corresponding data. *Tri-training* algorithm constitutes an improved single-view extension of the Co-training algorithm. This algorithm can be considered as a bagging ensemble of three classifiers which are trained on data subsets generated through bootstrap sampling from the original labeled training set (Hady & Schwenker 2010). In each Tri-training round, the

labeled set of each classifier is augmented with a unlabeled instance, labeled from the other two classifiers in case it disagrees. *Co-Bagging* algorithm (Hady & Schwenker 2010) creates several base classifiers using the same learning algorithm on a bootstrap sample created by random resampling with replacement from the original training set. Each bootstrap sample contains about 2/3 of the original training set, where each example can appear multiple times.

A high-level description of the DTCo algorithm is presented in Algorithm 1 which consists of two phases: *Training* phase and *Voting-Fusion* phase.

In the Training phase, the self-labeled algorithms which constitute the ensemble i.e., Democratic Co-learning, Tri-training and Co-Bagging, are independently trained using the same labeled L and unlabeled U datasets (Steps 1-3). In the Voting-Fusion phase, the trained self-labeled algorithms are applied on each instance in the testing set (Step 5). Next, the individual predictions of the three algorithms are combined via a maximum probability-based voting scheme. More specifically, the self-labeled algorithm which exhibits the most confident prediction over an unlabeled example of the test set is selected (Step 7). In case the confidence of the prediction of the selected classifier meets a predefined threshold (*ThresLev*) then the classifier labels the example otherwise the prediction is not considered reliable enough (Step 9). In this case, the output of the ensemble is defined as the combined predictions of three self-labeled learning algorithms via a simple majority voting, namely the ensemble output is the one made by more than half of them (Step 11).

Algorithm 1 DTCo

Input: L − Set of labeled instances.
U − Set of unlabeled instances.
C − Base learner.

Output: The labels of instances in the testing set.

/* Phase I: Training */
1: Democratic Co(L, U)
2: Tri-training(L, U)
3: Co-Bagging(L, U)

/* Phase II: Voting-Fusion */
4: **for each** $x \in T$ **do**
5: Apply Democratic Co-learning, Tri-training and Co-Bagging on x.
6: Find the classifier C^* with the highest confidence prediction on x.
7: **if** (Confidence of $C^* \geq$ *ThresLev*) **then**
8: C^* predicts the label y of x.
9: **else**
10: Use majority vote to predict the label y of x.
11: **end if**
12: **end for**

4 Experimental methodology

In this section, we conducted a of experiments in order to evaluate the performance of the proposed algorithm DTCo against the most popular self-labeled algorithms i.e. Self-training (Yarowsky 1995), Co-training (Blum & Mitchell 1998), Tri-training (Zhou & Li 2005), Co-Bagging (Hady & Schwenker 2010), CST-Voting (Livieris et al. 2018), Co-Forest (Li & Zhou 2007) and Democratic-Co learning (Zhou & Goldman 2004). The first five self-labeled methods were evaluated by deploying as base learners the MultiLayer Perceptron (MLP), the kNN algorithm and the $C4.5$ decision tree algorithm as in Livieris et al. (2018), Triguero et al. (2015).

The implementation code was written in Java, using the WEKA 3.9 Machine Learning Toolkit (Hall et al. 2009), In order to study the influence of the amount of labeled data, four different ratios (R) of the training data were used, i.e., 10%, 20%, 30% and 40%. All self-labeled algorithms utilized the configuration parameter settings as in Triguero et al. (2015) while all base learners were used with their default parameter settings included in the WEKA 3.9 library. Moreover, similar to Livieris (2019), we set parameter $ThresLev = 95\%$.

4.1 Datasets

The classification algorithms were evaluated using the the Shenzhen lung mask (Tuberculosis) dataset and the chest X-ray (Pneumonia) dataset.

Shenzhen lung mask (Tuberculosis) dataset: The dataset[1] was constructed by manually-segmented lung masks for the Shenzhen Hospital X-ray set as presented in Stirenko et al. (2018). These segmented lung masks were original utilized for the description of the lung segmentation technique in combination with lossless and lossy data augmentation. The segmentation masks for the Shenzhen Hospital X-ray set were manually prepared by students and teachers of the Computer Engineering Department, Faculty of Informatics and Computer Engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" Jaeger et al. (2014). The set contained 326 normal CXRs and 336 abnormal ones with tuberculosis, collected within a one-month period, mostly in September 2012. Notice that all algorithms were evaluated using the stratified 10-fold cross-validation on this dataset.

Chest X-ray (Pneumonia) dataset: The dataset[2] contains 5830 chest X-ray images (anterior-posterior), selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou. All chest X-ray imaging was performed as part of patients' routine clinical care. For the analysis of chest X-ray images, all chest radiographs were initially screened for quality control by removing all low quality or unreadable scans. The diagnoses for the images were then graded by two expert physicians and in order to account for any grading errors, the evaluation set was also checked by a third expert. Similar to Kermany et al. (2018), the dataset was partitioned into two sets (training/testing). The training set consisting of 5216 examples (1341 normal, 3875 pneumonia) and the testing set with 624 examples (234 normal, 390 pneumonia).

4.2 Performance evaluation of self-labeled algorithms

The performance of the self-labeled algorithms was evaluated using the following four performance metrics: Sensitivity (Sen), Specificity (Spe), F-measure (F_1) and Accuracy (Acc) which are respectively defined by

$$Sen = \frac{T_P}{T_P + F_N}, \qquad Spe = \frac{T_N}{T_N + F_P}, \qquad F_1 = \frac{2T_P}{2T_P + F_N + F_P} \qquad Acc = \frac{T_P + T_N}{T_P + T_N + F_P + F_N},$$

where T_P stands for the number of instances which have been correctly classified as positive, T_N stands for the number of instances which have been correctly classified as negative, F_P (type I error) stands for the number of instances which have been wrongly classified as positive, F_N (type II error) stands for the number of instances which have been wrongly classified as negative. Sensitivity of classification is the proportion of actual positives that are predicted as positive; Specificity represents the proportion of actual negatives that are predicted as negative; F_1 consists of a harmonic mean of precision and recall; while Accuracy is the ratio of correct predictions of a classifier.

[1] `https://www.kaggle.com/kmader/pulmonary-chest-xray-abnormalities/home`

[2] `https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia`

Tables 1 and 2 present the performance of all self-labeled methods for Tuberculosis dataset using labeled ratio $10\% - 20\%$ and $30\% - 40\%$, respectively. Notice that the highest classification performance for labeled ratio and performance metric is highlighted in bold. The aggregated results showed that DTCo was the most efficient and robust method independent of the utilized ratio of labeled instances in the training set.

Algorithm	Ratio = 10%				Ratio = 20%			
	Sen	Spe	F_1	Acc	Sen	Spe	F_1	Acc
Self-train (MLP)	65.23%	64.46%	64.65%	64.84%	67.74%	66.20%	66.90%	66.96%
Self-train (C4.5)	63.08%	69.69%	64.94%	66.43%	67.74%	67.25%	67.26%	67.49%
Self-train (NB)	68.82%	62.37%	66.32%	65.55%	67.74%	63.76%	66.08%	65.72%
Co-train (MLP)	62.72%	64.46%	62.95%	63.60%	64.52%	64.81%	64.29%	64.66%
Co-train (C4.5)	70.97%	54.01%	65.02%	62.37%	70.61%	58.54%	66.22%	64.49%
Co-train (NB)	58.06%	65.16%	59.89%	61.66%	65.23%	67.60%	65.70%	66.43%
Tri-train (MLP)	67.38%	65.51%	66.43%	66.43%	67.03%	67.94%	67.03%	67.49%
Tri-train (C4.5)	64.16%	**70.03%**	65.81%	67.14%	67.38%	66.55%	66.79%	66.96%
Tri-train (NB)	65.23%	66.20%	65.23%	65.72%	64.87%	63.07%	63.96%	63.96%
CST-Voting (MLP)	67.74%	64.81%	66.43%	66.25%	67.38%	67.60%	67.14%	67.49%
CST-Voting (C4.5)	66.31%	68.29%	66.67%	67.31%	69.53%	66.55%	68.19%	68.02%
CST-Voting (NB)	65.95%	66.90%	65.95%	66.43%	66.31%	67.60%	66.43%	66.96%
Co-Bagging (MLP)	65.59%	66.20%	65.47%	65.90%	65.95%	66.20%	65.71%	66.08%
Co-Bagging (C4.5)	64.87%	60.63%	63.18%	62.72%	66.31%	62.72%	64.80%	64.49%
Co-Bagging (NB)	64.52%	66.20%	64.75%	65.37%	65.23%	66.90%	65.47%	66.08%
Co-Forest	62.01%	58.89%	60.70%	60.42%	66.31%	59.58%	63.79%	62.90%
Democratic-Co	**71.33%**	67.60%	69.70%	69.43%	**71.68%**	67.25%	69.81%	69.43%
DTCo	70.61%	68.99%	**69.73%**	**69.79%**	70.97%	**68.99%**	**69.96%**	**69.96%**

Table 1 Performance of all self-labeled algorithms for ratio $R = 10\%$ and $R = 20\%$ for Tuberculosis dataset

Algorithm	Ratio = 30%				Ratio = 40%			
	Sen	Spe	F_1	Acc	Sen	Spe	F_1	Acc
Self-train (MLP)	68.10%	66.90%	67.38%	67.49%	65.23%	68.99%	66.18%	67.14%
Self-train (C4.5)	62.01%	68.64%	63.84%	65.37%	62.72%	68.64%	64.34%	65.72%
Self-train (NB)	66.31%	68.64%	66.79%	67.49%	68.10%	68.64%	67.98%	68.37%
Co-train (C4.5)	69.89%	65.85%	68.18%	67.84%	70.25%	67.60%	69.01%	68.90%
Co-train (MLP)	65.23%	66.20%	65.23%	65.72%	65.23%	66.55%	65.35%	65.90%
Co-train (NB)	68.82%	67.94%	68.21%	68.37%	68.82%	67.94%	68.21%	68.37%
Tri-train (MLP)	68.10%	69.34%	68.22%	68.73%	68.10%	69.69%	68.35%	68.90%
Tri-train (C4.5)	64.52%	69.34%	65.81%	66.96%	68.10%	69.34%	68.22%	68.73%
Tri-train (NB)	65.95%	66.55%	65.83%	66.25%	66.67%	66.90%	66.43%	66.78%
CST-Voting (MLP)	68.10%	68.64%	67.98%	68.37%	65.95%	69.69%	66.91%	67.84%
CST-Voting (C4.5)	67.03%	69.34%	67.51%	68.20%	69.89%	69.69%	69.52%	69.79%
CST-Voting (NB)	68.46%	68.64%	68.21%	68.55%	69.18%	69.34%	68.93%	69.26%
Co-Bagging (MLP)	65.95%	66.90%	65.95%	66.43%	68.46%	68.29%	68.09%	68.37%
Co-Bagging (C4.5)	71.68%	66.55%	69.57%	69.08%	72.76%	**70.03%**	71.48%	71.38%
Co-Bagging (NB)	65.95%	66.90%	65.95%	66.43%	66.31%	67.94%	66.55%	67.14%
Co-Forest	67.03%	66.20%	66.43%	66.61%	68.10%	66.20%	67.14%	67.14%
Democratic-Co	72.40%	66.90%	70.14%	69.61%	66.31%	68.99%	66.91%	67.67%
DTCo	**73.12%**	**69.34%**	**71.45%**	**71.20%**	**73.48%**	69.69%	**71.80%**	**71.55%**

Table 2 Performance of all self-labeled algorithms for ratio $R = 30\%$ and $R = 40\%$ for Tuberculosis dataset

Tables 3 and 4 present the classification performance for the Pneumonia dataset using labeled ratio $10\% - 20\%$ and $30\% - 40\%$, respectively. As mentioned above, the accuracy measure of the best-performing algorithm is highlighted in bold. Similar observations can be made with the previous benchmark. More specifically, DTCo exhibited the best overall classification performance for *Acc* and F_1 performance metrics., relative to all utilized labeled ratio.

Algorithm	Ratio = 10%				Ratio = 20%			
	Sen	Spe	F_1	Acc	Sen	Spe	F_1	Acc
Self-train (MLP)	95.64%	47.44%	84.20%	77.56%	97.95%	33.33%	82.33%	73.72%
Self-train (C4.5)	93.59%	53.42%	84.49%	78.53%	93.85%	53.85%	84.72%	78.85%
Self-train (NB)	93.85%	44.87%	82.71%	75.48%	94.36%	45.30%	83.07%	75.96%
Co-train (C4.5)	96.15%	44.02%	83.71%	76.60%	96.67%	44.44%	84.06%	77.08%
Co-train (MLP)	**97.18%**	38.46%	83.02%	75.16%	97.69%	34.19%	82.38%	73.88%
Co-train (NB)	96.92%	32.05%	81.55%	72.60%	96.92%	32.05%	81.55%	72.60%
Tri-train (MLP)	96.15%	43.59%	83.61%	76.44%	94.87%	45.73%	83.43%	76.44%
Tri-train (C4.5)	93.59%	**57.26%**	85.38%	79.97%	94.10%	**57.69%**	85.75%	80.45%
Tri-train (NB)	93.59%	44.44%	82.49%	75.16%	92.82%	44.44%	82.09%	74.68%
CST-Voting (MLP)	96.92%	46.58%	84.66%	78.04%	97.69%	38.46%	83.28%	75.48%
CST-Voting (C4.5)	94.62%	55.56%	85.52%	79.97%	94.87%	56.84%	85.95%	80.61%
CST-Voting (NB)	95.13%	42.74%	82.91%	75.48%	95.13%	43.59%	83.09%	75.80%
Co-Bagging (MLP)	45.64%	45.30%	51.15%	45.51%	96.15%	37.61%	82.33%	74.20%
Co-Bagging (C4.5)	92.56%	53.85%	84.05%	78.04%	93.59%	56.84%	85.28%	79.81%
Co-Bagging (NB)	90.77%	47.44%	81.66%	74.52%	91.54%	50.85%	82.83%	76.28%
Co-Forest	97.18%	27.35%	80.72%	70.99%	**98.46%**	35.04%	82.94%	74.68%
Democratic-Co	96.15%	47.01%	84.36%	77.72%	97.18%	47.44%	84.98%	78.53%
DTCo	97.69%	54.27%	**86.79%**	**81.41%**	97.95%	55.98%	**87.31%**	**82.21%**

Table 3 Performance of all self-labeled algorithms for ratio $R = 10\%$ and $R = 20\%$ for Prnumonia dataset

Algorithm	Ratio = 30%				Ratio = 40%			
	Sen	Spe	F_1	Acc	Sen	Spe	F_1	Acc
Self-train (MLP)	**98.46%**	34.19%	82.76%	74.36%	97.18%	40.60%	83.48%	75.96%
Self-train (NB)	93.85%	45.30%	82.81%	75.64%	94.62%	46.15%	83.39%	76.44%
Self-train (C4.5)	94.10%	56.84%	85.55%	80.13%	94.10%	57.26%	85.65%	80.29%
Co-train (MLP)	97.18%	38.89%	83.11%	75.32%	98.21%	37.18%	83.26%	75.32%
Co-train (C4.5)	96.67%	44.44%	84.06%	77.08%	96.92%	44.44%	84.19%	77.24%
Co-train (NB)	96.92%	32.05%	81.55%	72.60%	96.92%	32.91%	81.73%	72.92%
Tri-train (MLP)	96.41%	47.01%	84.49%	77.88%	96.41%	47.86%	84.68%	78.21%
Tri-train (C4.5)	94.10%	58.12%	85.85%	80.61%	94.87%	58.55%	86.35%	81.25%
Tri-train (NB)	91.54%	45.30%	81.60%	74.20%	93.08%	47.01%	82.78%	75.80%
CST-Voting (MLP)	97.69%	38.89%	83.37%	75.64%	97.69%	41.45%	83.92%	76.60%
CST-Voting (C4.5)	95.13%	**59.40%**	86.68%	81.73%	95.13%	**59.83%**	86.78%	81.89%
CST-Voting (NB)	94.87%	44.44%	83.15%	75.96%	95.64%	45.30%	83.73%	76.76%
Co-Bagging (MLP)	96.67%	43.16%	83.78%	76.60%	96.67%	44.87%	84.15%	77.24%
Co-Bagging (C4.5)	94.10%	57.69%	85.75%	80.45%	95.13%	57.69%	86.28%	81.09%
Co-Bagging (NB)	92.31%	51.28%	83.33%	76.92%	91.79%	51.71%	83.16%	76.76%
Co-Forest	98.21%	41.03%	84.08%	76.76%	97.69%	40.17%	83.64%	76.12%
Democratic-Co	97.69%	47.44%	85.23%	78.85%	**98.21%**	51.71%	86.46%	80.77%
DTCo	97.95%	57.26%	**87.61%**	**82.69%**	**98.21%**	55.56%	**87.34%**	**82.21%**

Table 4 Performance of all self-labeled algorithms for ratio $R = 30\%$ and $R = 40\%$ for Prnumonia dataset

5 Conclusions

In this work, we proposed a new semi-supervised self-labeled algorithm for X-rays classification, called DTCo, based on an ensemble philosophy. DTCo combines the individual predictions of three efficient self-labeled algorithms, i.e., i.e., Democratic Co-learning, Tri-training and Co-Bagging, endorsing a maximum-probability voting process. The efficacy of the proposed algorithm was demonstrated by a series of experiments on two chest X-ray datasets, illustrating that reliable and robust classification models could be developed by the adaptation of ensemble methodologies in the semi-supervised learning framework.

Our future work is focused on enhancing the classification efficiency of DTCo utilizing more efficient and sophisticated ensemble schemes for combination of predictions of the self-labeled algorithms. Furthermore, since our numerical experiments are quite promising another interesting aspect is focusing on expanding our experiments and applying further the proposed algorithm to several biomedical datasets for image classification.

References

Alam, J., Alam, S. & Hossan, A. (2018), Multi-stage lung cancer detection and prediction using multi-class svm classifie, *in* '2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering', IEEE, pp. 1–4.

Blum, A. & Mitchell, T. (1998), Combining labeled and unlabeled data with co-training, *in* '11th annual conference on computational learning theory', pp. 92–100.

Candemir, S., Jaeger, S., Musco, K. P. J., Singh, R., Xue, Z., Karargyris, A., Antani, S., Thoma, G. & McDonald, C. (2014), 'Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration', *IEEE transactions on medical imaging* **33**, 577–590.

Hady, M. & Schwenker, F. (2010), 'Combining committee-based semi-supervised learning and active learning', *Journal of Computer Science and Technology* **25**(4), 681–698.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. (2009), 'The WEKA data mining software: An update', *SIGKDD explorations newsletters* **11**, 10–18.

Hogeweg, L., Mol, C., de Jong, P., Ayles, R. & van Ginneken, B. (2010), Fusion of local and global detection systems to detect tuberculosis in chest radiographs, *in* 'Medical image computing and computer-assisted intervention', pp. 650–657.

Hogeweg, L., Sánchez, C., de Jong, P., Maduskar, P. & van Ginneken, B. (2012), 'Clavicle segmentation in chest radiographs', *Medical image analysis* **16**(8), 1490–1502.

Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., Xue, Z., Palaniappan, K., Singh, R., Antani, S., Thoma, G., Wang, Y., Lu, P. & McDonald, C. (2014), 'Automatic tuberculosis screening using chest radiographs', *IEEE transactions on medical imaging* **33**, 233–245.

Kermany, D., Goldbaum, M., Cai, W., Valentim, C., Liang, H., Baxter, S., McKeown, A., Yang, G., Wu, X. & Yan, F. (2018), 'Identifying medical diagnoses and treatable diseases by image-based deep learning', *Cell* **172**(5), 1122–1131.

Kononenko, I. (2001), 'Machine learning for medical diagnosis: history, state of the art and perspective', *Artificial Intelligence in medicine* **23**(1), 89–109.

Li, M. & Zhou, Z. (2007), 'Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples', *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* **37**(6), 1088–1098.

Livieris, I. E. (2019), 'A new ensemble self-labeled semi-supervised algorithm', *Informatica* pp. 1–14, (to be appear).

Livieris, I. E., Kanavos, A., Tampakas, V. & Pintelas, P. (2018), 'An ensemble SSL algorithm for efficient chest X-ray image classification', *Journal of Imaging* **4**(7).

Muyoyeta, M., Maduskar, P., Moyo, M., Kasese, N., Milimo, D., Spooner, R., Kapata, N., Hogeweg, L., van Ginneken, B. & Ayles, H. (2014), 'The sensitivity and specificity of using a computer aided diagnosis program for automatically scoring chest X-rays of presumptive TB patients compared with xpert mtb/rif in lusaka zambia', *PloS one* **9**(4), e93757.

Plankis, T., Juozapavicius, A., Stasiene, E. & Usonis, V. (2017), 'Computer-aided detection of interstitial lung diseases: A texture approach', *Nonlinear Analysis* **22**(3), 404–411.

Santosh, K. & Antani, S. (2018), 'Automated chest X-ray screening: Can lung region symmetry help detect pulmonary abnormalities?', *IEEE transactions on medical imaging* **37**(5), 1168–1177.

Stirenko, S., Kochura, Y., Alienin, O., Rokovyi, O., Gang, P., Zeng, W. & Gordienko, Y. (2018), 'Chest X-ray analysis of tuberculosis by deep learning with segmentation and augmentation', *arXiv preprint arXiv:1803.01199* .

Triguero, I., García, S. & Herrera, F. (2015), 'Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study', *Knowledge and information systems* **42**(2), 245–284.

van Ginneken, B., Hogeweg, L. & Prokop, M. (2009), 'Computer-aided diagnosis in chest radiography: Beyond nodules', *European journal of radiology* **72**(2), 226–230.

World Health Organization (2017), 'Global tuberculosis report 2017'.

Xu, T., Cheng, I. & Mandal, M. (2011), Automated cavity detection of infectious pulmonary tuberculosis in chest radiographs, *in* 'IEEE international conference on engineering in medicine and biology society', pp. 5178–5181.

Yarowsky, D. (1995), Unsupervised word sense disambiguation rivaling supervised methods, *in* 'Proceedings of the 33rd annual meeting of the association for computational linguistics', pp. 189–196.

Zhou, Y. & Goldman, S. (2004), Democratic co-learning, *in* '16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)', IEEE, pp. 594–602.

Zhou, Z. (2011), When semi-supervised learning meets ensemble learning, Vol. 6, Springer, pp. 6–16.

Zhou, Z. & Li, M. (2005), 'Tri-training: Exploiting unlabeled data using three classifiers', *IEEE transactions on knowledge and data engineering* **17**(11), 1529–1541.

Zhu, X. (2011), Semi-supervised learning, *in* 'Encyclopedia of machine learning', Springer, pp. 892–897.

Zhu, X. & Goldberg, A. (2009), 'Introduction to semi-supervised learning', *Synthesis lectures on artificial intelligence and machine learning* **3**(1), 1–130.