Available online at www.sciencedirect.com

ScienceDirect

Electronic Notes in Theoretical Computer Science 343 (2019) 19–33

Electronic Notes in Theoretical Computer Science

www.elsevier.com/locate/entcs

Detecting Lung Abnormalities From X-rays Using an Improved SSL Algorithm

Ioannis Livieris[a,1,2] Andreas Kanavos[b,3] Panagiotis Pintelas[a,4]

[a] *Department of Mathematics,*
University of Patras,
Patras, Greece.

[b] *Department of Computer Engineering & Informatics*
University of Patras,
Patras, Greece.

Abstract

A significant component in computer-aided medical diagnosis is the automatic detection of lung abnormalities from digital chest X-rays; thus it constitutes a vital first step in radiologic image analysis. During the last decades, the rapid advances of digital technology and chest radiography have ultimately led to the development of large repositories with labeled and unlabeled images. Semi-supervised learning algorithms have become a hot topic of research, exploiting the explicit classification information of labeled images with the knowledge hidden in the unlabeled images. In the present work, we propose a new semi-supervised learning algorithm for the classification of lung abnormalities from X-rays based on an ensemble philosophy. The efficacy of the presented algorithm is demonstrated by numerical experiments, illustrating that reliable prediction models could be developed by incorporating ensemble methodologies in the semi-supervised framework.

Keywords: Semi-supervised learning; self-labeled algorithms; ensemble learning; majority voting, image classification.

1 Introduction

Despite the advances in medicine, as well as the development of efficient treatments, the diseases caused by lung abnormalities are considered to be of the greatest lethal diseases worldwide. According to the World Health Organization (WHO), pneumonia kills about 1.5 million children under 5 years old every year and only in 2013, it

[1] Corresponding author

[2] Email:livieris@teiwest.gr

[3] Email:kanavos@ceid.upatras.gr

[4] Email:ppintelas@gmail.com

https://doi.org/10.1016/j.entcs.2019.04.008

was estimated that 1.5 million people died of tuberculosis and 9 million new cases occurred [26].

A typical method for the detection of lung abnormalities consists of a posterior-anterior Chest X-Ray (CXR) in order to search the lung region for any abnormalities that could be present. Due to its easy accessibility and relatively low cost, CXR imaging is widely used for health diagnosis and monitoring. In medical centers, the image interpretation has been mostly performed by human experts and it is considered a long and complicated process. Nevertheless, distinguishing the various chest pathologies is a difficult and challenging task, even to the expert human observer. As a result, the area of diagnostic medicine has massively changed; from a rather qualitative science that was based on observations of whole organisms to a more quantitative science, which is also based on knowledge extraction from databases [24]. More specifically, research focused on developing intelligent computer-aided diagnosis systems for the automatic recognition of abnormalities from CXRs in order to assist radiologists in identifying and integrating all useful information available in a chest image. These systems incorporate machine learning and data mining techniques in order to exploit vast amount of information provided by patients' records and laboratory data (see [9, 11, 17, 27, 33, 34, 38] and the references therein). Along this line, several methodologies and techniques have been proposed, aiming at:

- classifying and/or detecting the presence of an abnormality (image classification);
- identifying the boundaries of lung for extracting quantitative information and segmenting images into normal and abnormal (medical image segmentation).

Mansoor et al. [25] presented an extended review and explained the capabilities and performance of currently available approaches for segmenting lungs with pathologic conditions on chest tomography images. Furthermore, they divided the lung field segmentation methods into five broad categories, with an overview of relative advantages and disadvantages of the methods belonging to each group.

Candermir et al. [6] proposed a robust lung segmentation method which detects lung boundaries utilizing image retrieval-based patient specific adaptive lung models. Their proposed 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. They presented a series of experiments utilizing 585 chest radiographs from three different datasets, which demonstrated the efficacy and robustness of the proposed approach. Rajaraman et al. [30] proposed a decision support system which is based on a convolutional neural network to expedite accurate diagnosis of the pathology. Their proposed system detects pneumonia in pediatric CXRs and further differentiates between bacterial and viral types to facilitate swift referrals which require urgent medical intervention.

In more recent works, Santosh and Antani [32] developed a novel concept which takes into account right and left lung region changes in terms of symmetry for detecting the evidence of tuberculosis. Their method utilizes common pulmonary abnormalities exhibited in CXR images including cavitations, consolidations, infiltrates, blunted costophrenic angles, opacities, pleural effusion. 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. Alam et al. [2] 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. Moreover, image scaling, color space transformation and contrast enhancement have been used for image enhancement while threshold and marker-controlled watershed based segmentation has been used for segmentation. Subsequently, a set of textural features extracted from the separated regions of interest is classified and the algorithm can efficiently detect whether the input image contains a tumor or not.

Nevertheless, despite all these efforts, there is still no widely-utilized method; mostly due to fact that the progress in the field has been hampered by the lack of available labeled images for efficiently training an accurate supervised classifier. With the rapid advances in digital chest radiography, the vigorous development of the Internet and the widespread adoption of electronic medical records, research centers have accumulated large repositories of classified (labeled) images and mostly of unclassified (unlabeled) images from human experts. By leveraging these images researchers and medical staff have a significant potential to transform biomedical research and the delivery of healthcare. However, the process of correctly labeling new unlabeled CXRs frequently requires the efforts of specialized personnel and expert physicians, which incurs high time and monetary costs.

To address this problem, Semi-Supervised Learning (SSL) algorithms constitute an appropriate machine learning methodology for extracting useful knowledge from both labeled and unlabeled data. The algorithms comprise characteristics of both supervised and unsupervised learning algorithms in order to efficiently combine the explicit classification information of labeled data with the hidden information in the unlabeled data in order to build efficient classifiers [7, 36]. 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, Zemmal et al. [41] implemented a computer assisted detection system for the diagnosis of breast cancer from mammographic images which is based on a semi-supervised support vector machine classifier. Along this line, Livieris et al. [24] proposed a semi-supervised learning algorithm for the classification of chest X-rays 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. Their numerical experiments presented the efficacy of the proposed SSL algorithm and its classification accuracy, therefore illustrating that reliable prediction models could be developed utilizing a few labeled and many unlabeled data.

Motivated by their work, we propose a new semi-supervised self-labeled algorithm, which is based on an ensemble philosophy. The proposed algorithm exploits the individual predictions of self-labeled algorithms, using a majority voting methodology. Our preliminary numerical experiments present the efficiency and the classification accuracy of the proposed algorithm, illustrating that reliable prediction models could be developed by incorporating ensemble methodologies in the semi-supervised framework.

The remainder of this paper is organized as follows: Section 2 presents a brief description of the semi-supervised self-labeled algorithms and Section 3 presents a detailed description of the proposed algorithm. Section 4 presents a series of experiments carried out in order to examine and evaluate the accuracy of the proposed algorithm against the most popular self-labeled classification algorithms. Finally, Section 5 discusses the conclusions and some research topics for future work.

2 On Semi-supervised Self-labeled Classification Algorithms

In this section, we present a formal definition of the semi-supervised classification problem. Let (x, y) be an example, where x belongs to a class y and a D-dimensional space in which x_i is the i-th attribute of the instance. Suppose that the training set $L \cup U$ consists of a labeled set L of N_L instances where y is known and of an unlabeled set U of N_U instances where y is unknown with $N_L \ll N_U$. Furthermore, there exists a test set T of N_T unseen instances where y is unknown, which has not been utilized in the training stage. It is worth noticing that the basic aim of the semi-supervised classification is to obtain an accurate learning hypothesis utilizing the training set, especially when the number of labeled instances is low.

Self-labeled methods constitute prominent SSL methods which address the shortage of labeled data via a self-learning process based on supervised prediction models. This class of algorithms is characterized by their simplicity of implementation as well as their wrapper-based philosophy. From a theoretical point of view, Triguero et al. [36] proposed an in-depth taxonomy based on the main characteristics presented in them and conducted an exhaustive study of their classification efficacy on several datasets. Next, we briefly describe the most relevant self-labeled approaches proposed in the literature which are divided into two main groups: Self-training and Co-training.

Self-training algorithm [40] is considered as one of the most simple and efficient algorithm to leverage unlabeled data. This algorithm wraps around a base learner and utilizes its own predictions to assign labels to unlabeled data. More analytically, a supervised classifier is initially trained on labeled examples and at each iteration the training is augmented gradually with classified unlabeled instances that have achieved a probability value over a defined threshold c. Nevertheless, in case noisy examples are characterized as confident, they can later be incorporated into the labeled training set; hence this technique can lead to erroneous predictions and low classification accuracy [44]. On the other hand, in standard *Co-training* algorithm [5], the attributes of data are split into two conditionally independent views. Subsequently, two classifiers are trained independently in each view and at each iteration they teach each other the most confidently predicted examples. Nigam and Ghani performed an extensive experimental analysis and concluded that the Co-training algorithm outperforms other self-labeled algorithms when a natural existence of two distinct and independent views exists. Unfortunately, the assumption about the existence of sufficient and redundant views is a luxury hardly met in most real-case scenarios. In general, the self-labeled algorithms proposed in the literature are based on the philosophy of these algorithms while most of them exploit on ensemble ideas and techniques.

The *Tri-training* algorithm [43] is probably the most representative approach, which is based on the ensemble philosophy, and constitutes an improved single-view extension of the Co-training algorithm. Generally, this algorithm attempts to determine the most reliable unlabeled data as the agreement of three classifiers and it 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 [13]. Specifically, in each Tri-training iteration, the labeled set of each classifier is augmented with a unlabeled instance, labeled from the other two classifiers in case it disagrees.

Democratic Co-learning algorithm [42] is based on the idea of ensemble learning and majority voting and follows the multi-view theory but from another aspect. More specifically, instead of demanding for multiple views of the corresponding data, it utilizes multiple algorithms for producing the necessary information and endorses a voted majority process for the final decision. Based on the previous works, Li and Zhou [18] proposed *Co-Forest* algorithm, which is based on the efficient training a number of Random trees on bootstrap data from the dataset. The basic idea of this algorithm is that the assignment of a few unlabeled examples to each Random tree during the training process. Eventually, the final decision is composed by a simple majority voting. It is worth noticing that the efficacy of Co-Forest is based on the utilization of Random trees, although the number of the available labeled examples is reduced. A rather similar approach was proposed by Hady and Schwenker [13] in which they proposed the *Co-Bagging* algorithm where confidence is estimated from the local accuracy of committee members. It 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.

In more recent works, Livieris et al. [24] and Livieris [20] proposed some ensemble self-labeled algorithms based on voting schemes. These algorithms combine the individual predictions of three self-labeled algorithms i.e. Self-training, Co-training and Tri-training utilizing a difference combination of voting mechanisms. Motivated by the previous works, in [23] the authors proposed a new semi-supervised learning algorithm, called AAST, which dynamically selects the most promising learner for a classification problem from a pool of classifiers based on a self-training philosophy. AAST initially uses several independent base learners and during the training process dynamically selects the most promising base learner relative to a strategy based on the number of the most confident predictions of unlabeled data.

3 A new Ensemble Self-labeled Algorithm

In this section, we present a detailed description of the proposed self-labeled algorithm, which is based on an ensemble philosophy, entitled EnSL (Ensemble Self-Labeled) algorithm.

Generally, the generation of an ensemble of classifiers considers mainly two steps: *Selection* and *Combination*. The selection of the component classifiers is considered essential for the efficiency of the ensemble while the key point for its efficacy is based on their diversity and their accuracy; while the combination of the individual classifiers' predictions takes place through several techniques with different philosophy [10, 31].

By taking these into consideration, the proposed algorithm is based on the idea of generating a set $C = (C_1, C_2, \ldots, C_N)$ of N self-labeled classifiers by applying different algorithms (with heterogeneous model representations) to a single dataset and the combination of their individual predictions takes place through a majority voting methodology.

A high-level description of the proposed algorithm, entitled EnSL, is presented in Algorithm 1 which consists of two phases: *Training* and *Voting-Fusion* phase. In the Training phase, the self-labeled algorithms, which constitute the ensemble are trained utilizing the same labeled L and unlabeled U datasets (Steps $1 - 3$). Next, in the Voting-Fusion phase, the final hypothesis on each unlabeled example x of the test set combines the individual predictions of self-labeled algorithms utilizing a majority voting methodology (Steps $4 - 9$). An overview of the proposed EnSL is depicted in Figure 1.

4 Experimental Methodology

In this section, we present a series of experiments in order to evaluate the performance of the proposed EnSL for X-ray classification against the most efficient and frequently utilized self-labeled algorithms. The implementation codes were written in Java, using the WEKA 3.9 Machine Learning Toolkit [14].

Our experimental results were obtained by conducting a two phase procedure: in the first phase, we evaluate the performance of the proposed algorithm EnSL against the most popular self-labeled algorithms, i.e. Self-training, Co-training, Tri-training, Co-Bagging, CST-Voting, Co-Forest and Democratic-Co learning, while in the second phase, we performed a statistical comparison between all compared semi-supervised self-labeled algorithms.

Algorithm 1 EnSL

```
Input:  $L$ – Set of labeled instances.
        $U$ – Set of unlabeled instances.
        $C = (C_1, C_2, \ldots, C_N)$ – Set of self-labeled classifiers which constitute the ensemble.

/* Phase I: Training */
for each $C_i \in C$ do
    Train $C_i$ using the labeled $L$ and the unlabeled dataset $U$.
end for

/* Phase II: Voting-Fusion */
for each $x \in T$ do
    for each $C_i \in C$ do
        Apply classifier $C_i$ on instance $x$.
    end for
    Use majority vote to predict the label $y^*$ of $x$.
end for

Output: The labels of instances in the testing set.
```

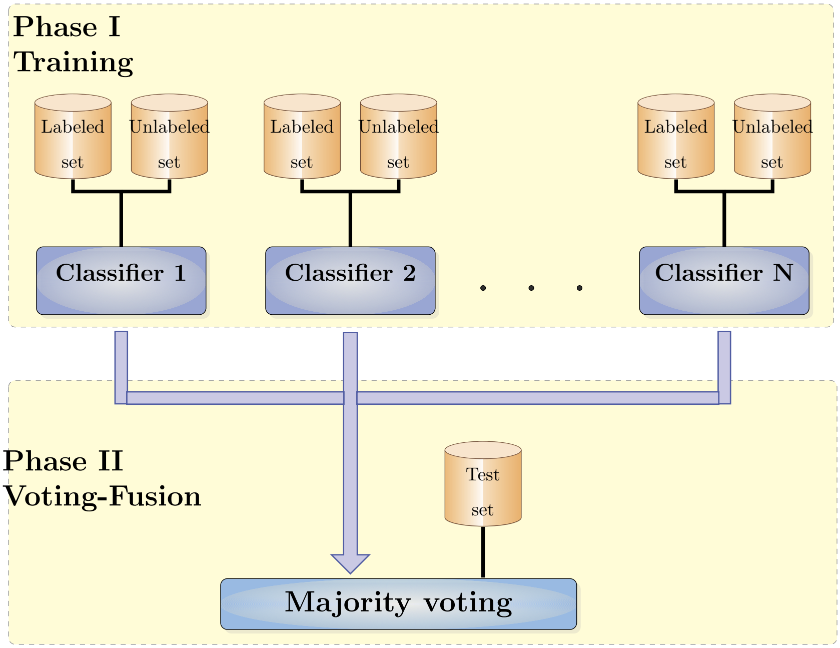

Fig. 1. EnSL algorithm

The performance of the classification algorithms is 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.

It is worth mentioning that 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.

4.1 Datasets

The compared semi-supervised learning classification algorithms were evaluated utilizing three different datasets: the chest X-ray (Pneumonia) as well as the CT Medical images dataset.

- *Chest X-ray (Pneumonia) dataset*: This dataset contains 5830 chest X-ray images (anterior-posterior) which were 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 before being cleared for training the artificial intelligence system. In order to account for any grading errors, the evaluation set was also checked by a third expert. Moreover, the dataset was partitioned into two sets (training/testing), where the training set consists of 5216 examples (1341 normal, 3875 pneumonia) and the testing set of 624 examples (234 normal, 390 pneumonia) as in [16].
- *CT Medical images dataset*: This data collection [1] contains 100 images [3] which constitute part of a much larger effort, focused on connecting cancer phenotypes to genotypes by providing clinical images matched to subjects from *the cancer genome Atlas* [8]. The images consist of the middle slice of all Computed Tomography (CT) images taken where valid age, modality and contrast tags could be found which results in 475 series from 69 different patients. Furthermore, this dataset is designed in order to allow different methods to be evaluated for examining the trends in CT image data associated with using contrast and patient age. The basic idea is to identify image textures, statistical patterns and

[1] https://www.kaggle.com/kmader/siim-medical-images/home

features correlating strongly with these traits and possibly build simple tools for automatically classifying these images when they have been misclassified (or finding outliers which could be suspicious cases, bad measurements, or poorly calibrated machines). Notice that all compared algorithms were evaluated using the stratified 10-fold cross-validation on this dataset.

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%.

4.2 Performance Evaluation of Semi-supervised Self-labeled Algorithms

In the sequel, we focus our interest on the experimental analysis for evaluating the classification performance of EnSL algorithm against its component self-labeled methods, i.e. Self-training, Co-training, Tri-training, Co-Bagging, CST-Voting, Co-Forest, Democratic-Co learning. Notice that the first five self-labeled methods were evaluated by deploying as base learners the Sequential Minimum Optimization (SMO) [28], the $C4.5$ decision tree algorithm [29] and the kNN algorithm [1]. These supervised classifiers probably constitute the most effective and popular machine learning algorithms for classification problems [39]. Moreover, similar to Blum and Mitchell [5], a limit to the number of iterations of all self-labeled algorithms is established which has also been adopted by many researchers [19–24, 35–37].

The self-labeled algorithms which constitute the ensemble of EnSL are: Self-training, Tri-training utilizing $C4.5$ as base learner, Co-training using (SMO) as base learner, Co-Forest and Democratic-Co learning. The motivation for this selection is based upon the fact that these algorithms have been reported as the most efficient self-labeled algorithms [36]. We recall that these methods are self-labeled ones, which exploit the hidden information in unlabeled data using different methodologies. More specifically, apart from the number of classifiers used by each method, the key concern is whether they are composed of the same (single) or different (multiple) learning algorithms. Self-training, Co-training, Tri-training and Co-Forest are single learning methods while Democratic-Co learning is a multiple learning method.

All self-labeled algorithms utilized the configuration parameter settings as in [4, 24,36]. Furthermore, similar to [20–24] all base learners were used with their default parameter settings included in the WEKA 3.9 library [14] in order to minimize the effect of any expert bias, instead of attempting to tune any of the algorithms to the specific dataset.

Tables 1 and 2 present the performance of all self-labeled methods on Pneumonia dataset, using labeled ratio equal to 10% − 20% and 30% − 40%, respectively. Notice that the highest classification performance for each labeled ratio and performance metric is highlighted in bold. The aggregated results showed that EnSL was the most efficient and robust method, independent of the utilized ratio of labeled instances in the training set. Moreover, it is worth noticing that EnSL exhibited the highest results for Acc and F_1 performance metrics.

Algorithm		Ratio = 10%				Ratio = 20%		
	Sen	Spe	F_1	Acc	Sen	Spe	F_1	Acc
Self-train (SMO)	95.13%	40.60%	82.44%	74.68%	95.90%	40.60%	82.83%	75.16%
Co-train (SMO)	94.10%	34.62%	80.66%	71.79%	94.36%	35.04%	80.88%	72.12%
Tri-train (SMO)	95.38%	39.32%	82.30%	74.36%	95.90%	40.17%	82.74%	75.00%
Co-Bagging (SMO)	94.87%	37.18%	81.59%	73.24%	95.90%	38.03%	82.29%	74.20%
CST-Voting (SMO)	96.92%	39.32%	83.08%	75.32%	96.92%	40.17%	83.26%	75.64%
Self-train (C4.5)	93.59%	53.42%	84.49%	78.53%	93.85%	53.85%	84.72%	78.85%
Co-train (C4.5)	96.15%	44.02%	83.71%	76.60%	96.67%	44.44%	84.06%	77.08%
Tri-train (C4.5)	93.59%	**57.26%**	85.38%	79.97%	94.10%	**57.69%**	85.75%	80.45%
Co-Bagging (C4.5)	92.56%	53.85%	84.05%	78.04%	93.59%	56.84%	85.28%	79.81%
CST-Voting (C4.5)	94.62%	55.56%	85.52%	79.97%	94.87%	56.84%	85.95%	80.61%
Self-train (kNN)	93.85%	44.87%	82.71%	75.48%	94.36%	45.30%	83.07%	75.96%
Co-train (kNN)	96.92%	32.05%	81.55%	72.60%	96.92%	32.05%	81.55%	72.60%
Tri-train (kNN)	93.59%	44.44%	82.49%	75.16%	92.82%	44.44%	82.09%	74.68%
Co-Bagging (kNN)	90.77%	47.44%	81.66%	74.52%	91.54%	50.85%	82.83%	76.28%
CST-Voting (kNN)	95.13%	42.74%	82.91%	75.48%	95.13%	43.59%	83.09%	75.80%
Co-Forest	97.18%	44.02%	84.22%	77.24%	**98.46%**	44.87%	85.05%	78.37%
Democratic-Co	96.15%	47.01%	84.36%	77.72%	97.18%	47.44%	84.98%	78.53%
EnSL	**97.95%**	55.98%	**87.31%**	**82.21%**	98.21%	54.70%	**87.14%**	**81.89%**

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

Algorithm		Ratio = 30%				Ratio = 40%		
	Sen	Spe	F_1	Acc	Sen	Spe	F_1	Acc
Self-train (SMO)	96.15%	40.60%	82.96%	75.32%	96.15%	40.60%	82.96%	75.32%
Co-train (SMO)	95.38%	35.90%	81.58%	73.08%	96.15%	37.18%	82.24%	74.04%
Tri-train (SMO)	95.90%	40.60%	82.83%	75.16%	95.90%	41.03%	82.93%	75.32%
Co-Bagging (SMO)	95.90%	39.32%	82.56%	74.68%	95.90%	42.74%	83.30%	75.96%
CST-Voting (SMO)	97.18%	40.17%	83.39%	75.80%	97.18%	40.17%	83.39%	75.80%
Self-train (C4.5)	94.10%	56.84%	85.55%	80.13%	94.10%	57.26%	85.65%	80.29%
Co-train (C4.5)	96.67%	44.44%	84.06%	77.08%	96.92%	44.44%	84.19%	77.24%
Tri-train (C4.5)	94.10%	58.12%	85.85%	80.61%	94.87%	58.55%	86.35%	81.25%
Co-Bagging (C4.5)	94.10%	57.69%	85.75%	80.45%	95.13%	57.69%	86.28%	81.09%
CST-Voting (C4.5)	95.13%	**59.40%**	86.68%	81.73%	95.13%	**59.83%**	86.78%	81.89%
Self-train (kNN)	93.85%	45.30%	82.81%	75.64%	94.62%	46.15%	83.39%	76.44%
Co-train (kNN)	96.92%	32.05%	81.55%	72.60%	96.92%	32.91%	81.73%	72.92%
Tri-train (kNN)	91.54%	45.30%	81.60%	74.20%	93.08%	47.01%	82.78%	75.80%
Co-Bagging (kNN)	92.31%	51.28%	83.33%	76.92%	91.79%	51.71%	83.16%	76.76%
CST-Voting (kNN)	94.87%	44.44%	83.15%	75.96%	95.64%	45.30%	83.73%	76.76%
Co-Forest	98.21%	45.73%	85.11%	78.53%	97.69%	46.15%	84.95%	78.37%
Democratic-Co	97.69%	47.44%	85.23%	78.85%	98.21%	51.71%	86.46%	80.77%
EnSL	**98.21%**	57.69%	**87.84%**	**83.01%**	**98.72%**	55.98%	**87.70%**	**82.69%**

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

Tables 3 and 4 present the performance of all self-labeled methods on CT Medical dataset, relative to all performance metrics, using labeled ratio equal to $10\% - 20\%$ and $30\% - 40\%$, respectively. As mentioned above, the accuracy measure of the best-performing self-labeled algorithm is highlighted in bold. Similar observations can be made as well with the previous benchmark. Firstly, it is worth mentioning that the proposed algorithm EnSL demonstrated the best performance. Regarding the F_1 and Acc metrics, EnSL exhibited the highest accuracy reporting the top performance in all cases, followed by CST-Voting. Finally, the results clearly show that EnSL increased its classification performance as the labeled ratio increased.

Algorithm		Ratio = 10%				Ratio = 20%		
	Sen	Spe	F_1	Acc	Sen	Spe	F_1	Acc
Self-train (SMO)	66.00%	62.00%	64.71%	64.00%	72.00%	68.00%	70.59%	70.00%
Co-train (SMO)	44.00%	64.00%	48.89%	54.00%	50.00%	**70.00%**	55.56%	60.00%
Tri-train (SMO)	66.00%	62.00%	64.71%	64.00%	72.00%	68.00%	70.59%	70.00%
Co-Bagging (SMO)	66.00%	62.00%	64.71%	64.00%	72.00%	68.00%	70.59%	70.00%
CST-Voting (SMO)	68.00%	64.00%	66.67%	66.00%	74.00%	**70.00%**	72.55%	72.00%
Self-train (C4.5)	64.00%	**66.00%**	64.65%	65.00%	68.00%	**70.00%**	68.69%	69.00%
Co-train (C4.5)	40.00%	50.00%	42.11%	45.00%	40.00%	54.00%	43.01%	47.00%
Tri-train (C4.5)	**72.00%**	62.00%	68.57%	67.00%	72.00%	66.00%	69.90%	69.00%
Co-Bagging (C4.5)	64.00%	64.00%	64.00%	64.00%	68.00%	68.00%	68.00%	68.00%
CST-Voting (C4.5)	70.00%	64.00%	67.96%	67.00%	74.00%	68.00%	71.84%	71.00%
Self-train (kNN)	62.00%	64.00%	62.63%	63.00%	68.00%	**70.00%**	68.69%	69.00%
Co-train (kNN)	34.00%	48.00%	36.56%	41.00%	40.00%	54.00%	43.01%	47.00%
Tri-train (kNN)	66.00%	62.00%	64.71%	64.00%	72.00%	66.00%	69.90%	69.00%
Co-Bagging (kNN)	48.00%	62.00%	51.61%	55.00%	54.00%	68.00%	58.06%	61.00%
CST-Voting (kNN)	66.00%	**66.00%**	66.00%	66.00%	72.00%	66.00%	69.90%	69.00%
Co-Forest	66.00%	60.00%	64.08%	63.00%	70.00%	60.00%	66.67%	65.00%
Democratic-Co	66.00%	62.00%	64.71%	64.00%	72.00%	68.00%	70.59%	70.00%
EnSL	**72.00%**	64.00%	**69.23%**	**68.00%**	**76.00%**	68.00%	**73.08%**	**72.00%**

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

4.3 Statistical and Post-Hoc Analysis

In machine learning, the statistical comparison of multiple algorithms over multiple datasets is fundamental, and usually it is carried out by means of a statistical test [22–24]. Since our motivation stems from the fact that we are interested in evaluating the rejection of the hypothesis that all the algorithms perform equally well for a given level based on their classification accuracy and highlighting the existence of significant differences between our proposed algorithm and the classical self-labeled algorithms, we utilized the non-parametric Friedman Aligned Ranking (FAR) [15] test. Furthermore, the Finner test [12] is applied as a post-hoc procedure in order to find out which algorithms present significant differences.

Table 5 presents the information of the statistical analysis performed by nonparametric multiple comparison procedures. The best (e.g. lowest) ranking obtained in each FAR test determines the control algorithm for the post-hoc test. Furthermore, the adjusted p-value with Finner's test (p_F) was presented based on the corresponding control algorithm, at the $\alpha = 0.05$ level of significance. It is worth mentioning that the test rejects the hypothesis of equality when the value of p_F is less than the

Algorithm	Ratio = 30%				Ratio = 40%			
	Sen	Spe	F_1	Acc	Sen	Spe	F_1	Acc
Self-train (SMO)	**78.00%**	68.00%	74.29%	73.00%	78.00%	70.00%	75.00%	74.00%
Co-train (SMO)	50.00%	70.00%	55.56%	60.00%	52.00%	70.00%	57.14%	61.00%
Tri-train (SMO)	76.00%	68.00%	73.08%	72.00%	78.00%	72.00%	75.73%	75.00%
Co-Bagging (SMO)	74.00%	68.00%	71.84%	71.00%	78.00%	70.00%	75.00%	74.00%
CST-Voting (SMO)	**78.00%**	68.00%	74.29%	73.00%	78.00%	70.00%	75.00%	74.00%
Self-train (C4.5)	74.00%	70.00%	72.55%	72.00%	78.00%	70.00%	75.00%	74.00%
Co-train (C4.5)	58.00%	54.00%	56.86%	56.00%	58.00%	60.00%	58.59%	59.00%
Tri-train (C4.5)	74.00%	66.00%	71.15%	70.00%	74.00%	68.00%	71.84%	71.00%
Co-Bagging (C4.5)	72.00%	70.00%	71.29%	71.00%	78.00%	72.00%	75.73%	75.00%
CST-Voting (C4.5)	76.00%	70.00%	73.79%	73.00%	76.00%	70.00%	73.79%	73.00%
Self-train (kNN)	72.00%	**72.00%**	72.00%	72.00%	74.00%	74.00%	74.00%	74.00%
Co-train (kNN)	58.00%	60.00%	58.59%	59.00%	60.00%	**76.00%**	65.22%	68.00%
Tri-train (kNN)	72.00%	66.00%	69.90%	69.00%	74.00%	66.00%	71.15%	70.00%
Co-Bagging (kNN)	66.00%	70.00%	67.35%	68.00%	72.00%	70.00%	71.29%	71.00%
CST-Voting (kNN)	76.00%	70.00%	73.79%	73.00%	78.00%	72.00%	75.73%	75.00%
Co-Forest	70.00%	64.00%	67.96%	67.00%	70.00%	66.00%	68.63%	68.00%
Democratic-Co	74.00%	68.00%	71.84%	71.00%	76.00%	68.00%	73.08%	72.00%
EnSL	**78.00%**	68.00%	**75.00%**	**74.00%**	**80.00%**	72.00%	**76.92%**	**76.00%**

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

value of a. Notice that, Self-training, Co-training, Tri-Training, Co-Bagging and CST-Voting utilized $C4.5$ as base learner, since they exhibited the best reported performance.

Clearly, EnSL demonstrates the best overall performance, as it outperforms the rest self-labeled algorithms. This is due to the fact that it reports the highest probability-based ranking by statistically presenting better results, relative to all labeled ratio.

Algorithm	FAR	Finner Post-Hoc Test	
		p_F-value	Null Hypothesis
EnSL	8.625	-	-
CST-Voting	17.875	0.120413	accepted
Tri-training	28.063	0.042806	rejected
Self-training	29.188	0.038961	rejected
Co-Bagging	29.875	0.038961	rejected
Democratic Co	34.625	0.012148	rejected
Co-Forest	51.75	0.000013	rejected
Co-training	60.0	0.000004	rejected

Table 5
Friedman Aligned Ranking (FAR) test and Finner Post-Hoc test

5 Conclusions

In this work, we proposed a new ensemble self-labeled algorithm for the detection of lung abnormalities from X-rays, entitled EnSL. The proposed algorithm combines the individual predictions of efficient self-labeled algorithms utilizing a majority voting methodology. For testing purposes, the algorithm was extensively evaluated on the chest X-rays (Pneumonia) dataset and the CT Medical images dataset utilizing Self-training, Co-training, Tri-training, Co-Bagging and Democratic-Co learning to constitute the ensemble. Our numerical experiments indicated the efficiency and the classification accuracy of the proposed algorithm EnSL, as statistically confirmed by the Friedman Aligned Ranks nonparametric test as well as the Finner post-hoc test. Therefore, we conclude that reliable and robust classification models could be developed by the adaptation of ensemble methodologies in the semi-supervised learning framework.

In our future work, we intend to pursue extensive empirical experiments for comparing the proposed EnSL with other algorithms, belonging to different SSL classes such as generative mixture models, transductive SVMs, as well as graph-based methods. Furthermore, since the experiments' results are quite encouraging, as a next step, we could consider the evaluation of the EnSL in several biomedical datasets for image classification as well as in specific scientific fields applying real-world datasets, such as educational, financial, healthcare, etc. and explore its performance on imbalanced datasets.

References

[1] D. Aha. *Lazy learning.* Dordrecht: Kluwer academic publishers, 1997.

[2] J. Alam, S. Alam, and A. Hossan. Multi-stage lung cancer detection and prediction using multi-class svm classifier. In *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering*, pages 1–4. IEEE, 2018.

[3] B. Albertina, M. Watson, C. Holback, R. Jarosz, S. Kirk, Y. Lee, and J. Lemmerman. Radiology data from the cancer genome atlas lung adenocarcinoma [TCGA-LUAD] collection. *The Cancer Imaging Archive*, 2016.

[4] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera. Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, 17, 2011.

[5] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *11th annual conference on computational learning theory*, pages 92–100, 1998.

[6] S. Candemir, S. Jaeger, K. Palaniappan J.P. Musco, R.K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma, and C.J. McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, 33:577–590, 2014.

[7] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-supervised learning.* MIT Press, Cambridge, MA, 2006.

[8] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, and M. Pringle. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, 2013.

[9] Mari Antonius Cornelis Dekker and Sandro Etalle. Audit-based access control for electronic health records. *Electronic Notes in Theoretical Computer Science*, 168:221–236, 2007.

[10] T.G. Dietterich. Ensemble methods in machine learning. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems*, volume 1857, pages 1–15. Springer Berlin Heidelberg, 2001.

[11] S. Dua, U.R. Acharya, and P. Dua. *Machine learning in healthcare informatics*, volume 56. Springer, 2014.

[12] H. Finner. On a monotonicity problem in step-down multiple test procedures. *Journal of the American statistical association*, 88(423):920–923, 1993.

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

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA data mining software: An update. *SIGKDD explorations newsletters*, 11:10–18, 2009.

[15] J.L. Hodges and E.L. Lehmann. Rank methods for combination of independent experiments in analysis of variance. *The annals of mathematical statistics*, 33(2):482–497, 1962.

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

[17] Evelina Lamma, L Maestrami, Paola Mello, Fabrizio Riguzzi, and Sergio Storari. Rule-based programming for building expert systems: A comparison in the microbiological data validation and surveillance domain. *Electronic Notes in Theoretical Computer Science*, 59(4):397–411, 2001.

[18] M. Li and Z.H. Zhou. 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, 2007.

[19] C. Liu and P.C. Yuen. A boosted co-training algorithm for human action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(9):1203–1213, 2011.

[20] I.E. Livieris. A new ensemble self-labeled semi-supervised algorithm. *Informatica*, (accepted for publication), 2018.

[21] I.E. Livieris, K. Drakopoulou, V. Tampakas, T. Mikropoulos, and P. Pintelas. An ensemble-based semi-supervised approach for predicting students' performance. In *Research on e-Learning and ICT in education*. Elsevier, 2018.

[22] I.E. Livieris, K. Drakopoulou, V. Tampakas, T. Mikropoulos, and P. Pintelas. Predicting secondary school students' performance utilizing a semi-supervised learning approach. *Journal of educational computing research*, 2018.

[23] I.E. Livieris, A. Kanavos, V. Tampakas, and P. Pintelas. An auto-adjustable semi-supervised self-training algorithm. *Algorithm*, 11(9), 2018.

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

[25] A. Mansoor, U. Bagci, B. Foster, Z. Xu, G.Z. Papadakis, L.R. Folio, J.K. Udupa, and D.J. Mollura. Segmentation and image analysis of abnormal lungs at CT: current approaches, challenges, and future trends. *RadioGraphics*, 35(4):1056–1076, 2015.

[26] World Health Organization. *Global tuberculosis report 2013.* World Health Organization, 2013.

[27] José Ramón Pasillas-Díaz and Sylvie Ratté. An unsupervised approach for combining scores of outlier detection techniques, based on similarity measures. *Electronic Notes in Theoretical Computer Science*, 329:61–77, 2016.

[28] J.C. Platt. *Advances in Kernel Methods - Support Vector Learning.* MIT Press, Cambridge, Massachusetts, 1998.

[29] J.R. Quinlan. *C4.5: Programs for machine learning.* Morgan Kaufmann, San Francisco, 1993.

[30] S. Rajaraman, S. Candemir, I. Kim, G. Thoma, and S. Antani. Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. *Applied Sciences*, 8(10):1715, 2018.

[31] L. Rokach. *Pattern classification using ensemble methods.* World Scientific Publishing Company, 2010.

[32] K.C. Santosh and S. Antani. Automated chest X-ray screening: Can lung region symmetry help detect pulmonary abnormalities? *IEEE Transactions on Medical Imaging*, 37(5):1168–1177, 2018.

[33] M. Sonka, V. Hlavac, and R. Boyle. *Image processing, analysis, and machine vision.* Cengage Learning, 2014.

[34] K. Suzuki. *Machine learning in computer-aided diagnosis: Medical imaging intelligence and analysis.* IGI Global, 2012.

[35] J. Tanha, M. van Someren, and H. Afsarmanesh. Semi-supervised self-training for decision tree classifiers. *International journal of machine learning cybernetics*, 8:355–370, 2015.

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

[37] I. Triguero, J.A. Sáez, J. Luengo, S. García, and F. Herrera. On the characterization of noise filters for self-training semi-supervised in nearest neighbor classification. *Neurocomputing*, 132:30–41, 2014.

[38] Ton van Deursen, Paul Koster, and Milan Petković. Hedaquin: A reputation-based health data quality indicator. *Electronic Notes in Theoretical Computer Science*, 197(2):159–167, 2008.

[39] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A.F.M. Ng, B. Liu, P.S. Yu, Z.H. Zhou, M. Steinbach, D.J. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.

[40] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995.

[41] N. Zemmal, N. Azizi, N. Dey, and M. Sellami. Adaptive semi supervised support vector machine semi supervised learning with features cooperation for breast cancer classification. *Journal of Medical Imaging and Health Informatics*, 6(1):53–62, 2016.

[42] Y. Zhou and S. Goldman. Democratic co-learning. In *16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 594–602. IEEE, 2004.

[43] Z.H. Zhou and M. Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1529–1541, 2005.

[44] X. Zhu and A.B. Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.