

Article

A Weighted Voting Ensemble Self-Labeled Algorithm for the Detection of Lung Abnormalities from X-Rays

Ioannis E. Livieris ^{1,*} , Andreas Kanavos ¹, Vassilis Tampakas ¹ and Panagiotis Pintelas ²

¹ Department of Computer & Informatics Engineering (DISK Lab), Technological Educational Institute of Western Greece, GR 263-34 Antirion, Greece; kanavos@ceid.upatras.gr (A.K.); vtampakas@teimes.gr (V.T.)

² Department of Mathematics, University of Patras, GR 265-00 Patras, Greece; ppintelas@gmail.com

* Correspondence: livieris@teiwest.gr

Received: 12 February 2019; Accepted: 13 March 2019; Published: 16 March 2019



Abstract: During the last decades, intensive efforts have been devoted to the extraction of useful knowledge from large volumes of medical data employing advanced machine learning and data mining techniques. Advances in digital chest radiography have enabled research and medical centers to accumulate large repositories of classified (labeled) images and mostly of unclassified (unlabeled) images from human experts. Machine learning methods such as semi-supervised learning algorithms have been proposed as a new direction to address the problem of shortage of available labeled data, by exploiting the explicit classification information of labeled data with the information hidden in the unlabeled data. In the present work, we propose a new ensemble semi-supervised learning algorithm for the classification of lung abnormalities from chest X-rays based on a new weighted voting scheme. The proposed algorithm assigns a vector of weights on each component classifier of the ensemble based on its accuracy on each class. Our numerical experiments illustrate the efficiency of the proposed ensemble methodology against other state-of-the-art classification methods.

Keywords: machine learning; semi-supervised learning; self-labeled algorithms; classifiers; ensemble learning; weighted voting; image classification; lung abnormalities

1. Introduction

The automatic detection of abnormalities, diseases and pathologies constitutes a significant factor in computer-aided medical diagnosis and a vital component in radiologic image analysis. For over a century, radiology has been a typical method for abnormality detection. A typical radiological examination is performed by utilizing a posterior–anterior chest radiograph, which is most commonly called Chest X-Ray (CXR). CXR imaging is widely used for health diagnosis and monitoring, due to its relatively low cost and easy accessibility; thus, it has been established as the single most acquired medical image modality [1]. It constitutes a significant factor for the detection and diagnosis of several pulmonary diseases, such as tuberculosis, lung cancer, pulmonary embolism and interstitial lung disease [1]. However, due to increasing workload pressures, many radiologists today have to daily examine an enormous number of CXRs. Thus, a prediction system trained to predict the risk of specific abnormalities given a particular CXR image is considered essential for providing high quality medical assistance. More specifically, such a decision support system has the potential to support the reading workflow, improve efficiency and reduce prediction errors. Moreover, it could be used to enhance the confidence of the radiologist or prioritize the reading list where critical cases would be read first.

The significant advances in digital chest radiography and the continuously enlarged 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. To this end, researchers and medical staff were able to leverage and exploit these images by the adoption

of machine learning and data mining techniques for the development of intelligent computational systems in order to extract useful and valuable information. As a result, the areas of biomedical research and diagnostic medicine have been dramatically transformed, from rather qualitative sciences which were based on observations of whole organisms to more quantitative sciences which are now based on the extraction of useful knowledge from a large amount of data [2].

Nevertheless, distinguishing the various chest abnormalities from CXRs is a rather challenging task, not only for a prediction model but even for an human expert. The progress in the field has been hampered by the lack of available labeled images for efficiently training a powerful and accurate supervised classification model. Moreover, the process of correctly labeling new unlabeled CXRs usually incurs monetary costs and high time since it constitutes a long and complicated process and requires the efforts of specialized personnel and expert physicians.

Semi-Supervised Learning (SSL) algorithms have been proposed as a new direction to address the problem of shortage of available labeled data, comprising characteristics of both supervised and unsupervised learning algorithms. These algorithms efficiently develop powerful classifiers by meaningfully relating the explicit classification information of labeled data with the information hidden in the unlabeled data [3,4]. Self-labeled algorithms probably constitute the most popular class of SSL algorithms due to their simplicity of implementation, their wrapper-based philosophy and good classification performance [2,5–8]. This class of algorithms exploits a large amount of unlabeled data via a self-learning process based on supervised learners. In other words, they perform an iterative procedure, enriching the initial labeled data, based on the assumption that their own predictions tend to be correct.

Recently, Triguero et al. [9] proposed an in-depth taxonomy based on the main characteristics presented in them and conducted a comprehensive research of their classification efficacy on several datasets. Generally, self-labeled algorithm can be classified in two main groups: *Self-training* and *Co-training*. In the original Self-training [10], a single classifier is iteratively trained on an enlarged labeled dataset with its most confident predictions on unlabeled data while in Co-training [11], two classifiers are separately trained utilizing two different views on a labeled dataset and then each classifier augments the labeled data of the other with its most confident predictions on unlabeled data. Along this line, several self-labeled algorithms have been proposed in the literature, while some of them exploit ensemble methodologies and techniques.

Democratic-Co learning [12] is based on an ensemble philosophy since it uses three independent classifiers following a majority voting and a confidence measurement strategy for predicting the values of unlabeled examples. Tri-training algorithm [13] utilizes a bagging ensemble of three classifiers which are trained on data subsets generated through bootstrap sampling from the original labeled set and teach each other using on majority voting strategy. Co-Forest [14] utilizes bootstrap sample data from the labeled set in order to train Random trees. At each iteration, each random tree is reconstructed by newly selected unlabeled instances for its concomitant ensemble, utilizing a majority voting technique. Co-Bagging [15] trains multiple base classifiers on bootstrap data created by random resampling with replacement from the training set. Each bootstrap sample contains about 2/3 of the original training set, where each example can appear multiple times. Recently, a new approach has been given by Livieris et al. [2,8,16,17] and Livieris [18] in which some ensemble self-labeled algorithms are proposed based on voting schemes. The proposed algorithms exploit the individual predictions of the most efficient and frequently used self-labeled algorithms using simple voting methodologies.

Motivated by these works, we propose a new semi-supervised self-labeled algorithm which is based on a sophisticated ensemble philosophy. The proposed algorithm exploits the individual predictions of self-labeled algorithms, using a new weighted voting methodology. The proposed weighted strategy assigns weights on each component classifier of the ensemble based on its accuracy on each class. Our main aim is to measure the effectiveness of our weighted voting ensemble scheme over the majority voting ensembles, using identical component classifiers in all cases. On top of that, we want to verify that powerful classification models could be developed by the adaptation of

advanced ensemble methodologies in the SSL framework. Our preliminary numerical experiments prove the efficiency and the classification accuracy of the proposed algorithm, demonstrating 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 survey of recent studies concerning the application of machine learning for the detection of lung abnormalities from X-rays. Section 3 presents a detailed description of the proposed weighted voting scheme and ensemble 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. Related Work

The significance of medical imaging for the diagnosis of diseases has been established for the treatment of chest pathologies and their early detection. During the last decades, the advances of digital technology and chest radiography as well as the rapid development of digital image retrieval have renewed the progress in new technologies for the diagnosis of lung abnormalities. More specifically, research has been focused on the development of Computer-Aided Diagnostic (CAD) models for abnormality detection in order to assist medical staff. Along this line, a variety of methodologies have been proposed based on machine learning techniques, aiming on classifying and/or detecting abnormalities in patients' medical images. A number of studies have been carried out in recent years; some useful outcomes of them are briefly presented below.

Jaeger et al. [19] proposed a CAD system for tuberculosis in conventional posteroanterior chest radiographs. Their proposed model initially utilizes a graph cut segmentation method to extract the lung region from the CXRs and then a set of texture and shape features in the lung region is computed in order to classify the patient as normal or abnormal. Their extensive numerical experiments on two real-world datasets illustrated the efficiency of the proposed CAD system for tuberculosis screening, achieving higher performance compared to that of human readings.

Melendez et al. [20] recommend a novel CAD system for detecting tuberculosis on chest X-rays based on multiple-instance learning. Their proposed system is based on the idea of utilizing probability estimations, instead of the sign of a decision function, to guide the multiple-instance learning process. Furthermore, an advantage of their method is that it does not require labeling of each feature sample during the training process but only a global class label characterizing a group of samples.

Alam et al. [21] utilized a multi-class support vector machine classifier and developed an efficient lung cancer detection and prediction model. The image enhancement and image segmentation have been done independently in every stage of the classification process. Image scaling, color space transformation and contrast enhancement have been utilized for image enhancement while threshold and marker-controlled watershed have been utilized for segmentation. In the sequel, the support vector machine classifier categorizes a set of textural features extracted from the separated regions of interest. Based on their numerical experiments, the authors concluded that the proposed algorithm can efficiently detect a cancer-affected cell and its corresponding stage such as initial, middle, or final. Furthermore, in case no cancer-affected cell is found in the input image then it checks the probability of lung cancer.

In more recent works, Madani [22] focused on the detection of abnormalities in chest X-ray images, having available only a fairly small size dataset of annotated images. Their proposed method deals with both problems of labeled data scarcity and data domain overfitting, by utilizing Generative Adversarial Networks (GAN) in a SSL architecture. In general, GAN utilize two networks: a generator which seeks to create as realistic images as possible and a discriminator which seeks to distinguish between real data and generated data. Next, these networks are involved in a minimax game to find the Nash equilibrium between them. Based on their experiments, the author concluded that the annotation effort is reduced considerably to achieve similar performance through supervised training techniques.

In [2], Livieris et al. evaluated the classification efficacy of an ensemble SSL algorithm, called CST-Voting, for CXR classification of tuberculosis. The proposed algorithm combines the individual predictions of three efficient self-labeled algorithms i.e., Co-training, Self-training and Tri-training using a simple majority voting methodology. The authors presented some interesting results, illustrating the efficiency of the proposed algorithm against several classical algorithms. Additionally, their experiments lead them to the conclusion that reliable and robust prediction models could be developed utilizing a few labeled and many unlabeled data. In [16] the authors extended the previous work and proposed DTCo algorithm for the classification of X-rays. The proposed ensemble algorithm exploits the predictions of Democratic-Co learning, Tri-training and Co-Bagging utilizing a maximum-probability voting scheme. Along this line, Livieris et al. [17] proposed EnSL algorithm which constitutes a generalized scheme of the previous works. More specifically, EnSL constitutes a majority voting scheme of N self-labeled algorithms. Their preliminary numerical experiments demonstrated that robust classification models could be developed by the adaptation of ensemble methodologies in the SSL framework.

Guan and Huang [23] considered the problem of multi-label thorax disease classification on chest X-ray images by proposing a Category-wise Residual Attention Learning (CRAL) framework. CRAL predicts the presence of multiple pathologies in a class-specific attentive view, aiming to suppress the obstacles of irrelevant classes by endowing small weights to the corresponding feature representation while the same time, the relevant features would be strengthened by assigning larger weights. More analytically, their proposed framework consists of two modules: feature embedding module and attention learning module. The feature embedding module learns high-level features using a neural network classifier while the attention learning module focuses on exploring the assignment scheme of different categories. Based on their numerical experiments, the authors stated that their proposed methodology constitutes a new state of the art.

3. A New Weighted Voting 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 Weighed voting Ensemble Self-Labeled (WvEnSL) 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 and 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 [24,25].

By taking these into consideration, the proposed algorithm is based on the idea of selecting a set $C = (C_1, C_2, \dots, C_N)$ 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 new weighted voting methodology. It is worth noticing that weighted voting is a commonly used strategy for combining predictions in pairwise classification in which the classifiers are not treated equally. Each classifier is evaluated on a evaluation set D and associated with a coefficient (weight), usually proportional to its classification accuracy.

Let us consider a dataset D with M classes, which is utilized for the evaluation of each component classifier. More specifically, the performance of each classifier C_i , with $i = 1, 2, \dots, N$ is evaluated on D and a $N \times M$ matrix W is defined, as follows

$$W = \begin{bmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,M} \\ w_{2,1} & w_{2,2} & \dots & w_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N,1} & w_{N,2} & \dots & w_{N,M} \end{bmatrix}$$

where each element $w_{i,j}$ is defined by

$$w_{i,j} = \frac{2p_j^{(C_i)}}{|D_j| + p_j^{(C_i)} + q_j^{(C_i)}}, \quad (1)$$

where D_j is the set of instances of the dataset belonging to the class j , $p_j^{(C_i)}$ are the number of correct predictions of classifier C_i on D_j and $q_j^{(C_i)}$ are the number of incorrect predictions of C_i that an instance belongs to class j . Clearly, each weight $w_{i,j}$ is the F_1 -score of classifier C_i for j class [26]. The rationale behind (1) is to measure the efficiency of each classifier, relative to each class j of the evaluation set D .

Subsequently, the class \hat{y} of each unknown instance x in the test set is computed by

$$\hat{y} = \arg \max_j \sum_{i=1}^N w_{i,j} \chi_A(C_i(x) = j),$$

where function $\arg \max$ returns the value of index corresponding to the largest value from array, $A = \{1, 2, \dots, M\}$ is the set of unique class labels and χ_A is the characteristic function which takes into account the prediction $j \in A$ of a classifier C_i on an instance x and creates a vector in which the j coordinate takes a value of one and the rest take the value of zero. At this point, it is worth mentioning that in our implementation we selected to evaluate the performance of each classifier of the ensemble on the initial training labeled set L .

A high-level description of the proposed framework is presented in Algorithm 1 which consists of three phases: *Training*, *Evaluation* and *Weighted-Voting Prediction*. In the Training phase, the self-labeled algorithms, which constitute the ensemble are trained utilizing the same labeled L and unlabeled dataset U (Steps 1–3). Subsequently, in the Evaluation phase, the trained classifiers are evaluated using the training set L in order to calculate the weight matrix W (Steps 4–9). Finally, in the Weighted-Voting Prediction phase, the final hypothesis on each unlabeled example x of the test set combines the individual predictions of self-labeled algorithms utilizing the proposed weighted voting methodology (Steps 10–15). An overview of the proposed WvEnSL is depicted in Figure 1.

Algorithm 1: WvEnSL

Input: L – Set of labeled instances (Training labeled set).
 U – Set of unlabeled instances (Training unlabeled set).
 T – Set of unlabeled test instances (Testing set).
 D – Set of instances for evaluation (Evaluation set).
 $C = (C_1, C_2, \dots, C_N)$ – Set of self-labeled classifiers which constitute the ensemble.

Output: The labels of instances in the testing set.

/* Phase I: Training */

Step 1: for $i = 1$ to N do

Step 2: Train C_i using the labeled L and the unlabeled dataset U .

Step 3: end for

/* Phase II: Evaluation */

Step 4: for $i = 1$ to N do

Step 5: Apply C_i on the evaluation set D .

Algorithm 1: *Cont.*

Step 6: for $j = 1$ to M do
Step 7: Calculate the weight

$$w_{i,j} = \frac{2p_j^{(C_i)}}{|D_j| + p_j^{(C_i)} + q_j^{(C_i)}}$$

Step 8: end for
Step 9: end for

/ Phase III: Weighted-Voting Prediction */*

Step 10: for each $x \in T$ do
Step 11: for $i = 1$ to N do
Step 12: Apply classifier C_i on x .
Step 13: end for
Step 14: Predict the label \hat{y} of x using

$$\hat{y} = \arg \max_j \sum_{i=1}^N w_{i,j} \chi_A(C_i(x) = j).$$

Step 15: end for

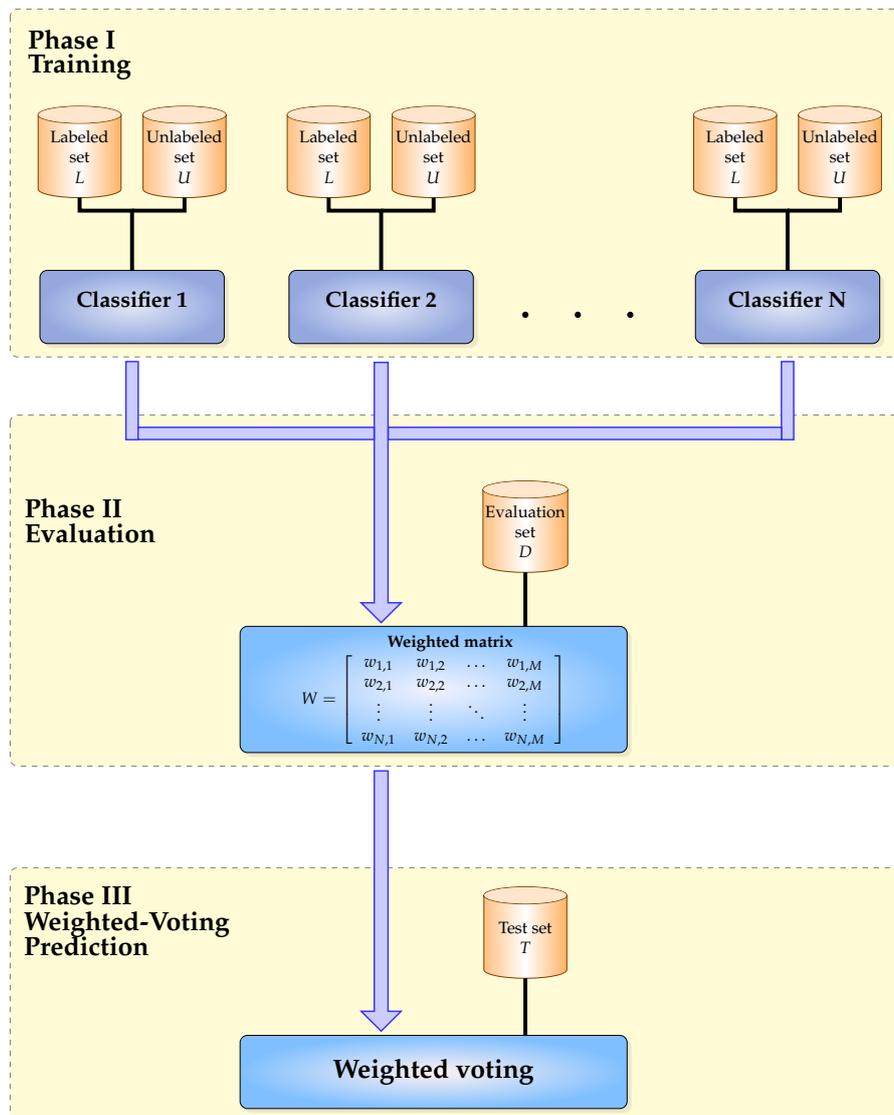


Figure 1. WvEnSL framework.

4. Experimental Methodology

In this section, we present a series of experiments in order to evaluate the performance of the proposed WvEnSL algorithm for X-ray classification against the most efficient ensemble self-labeled algorithms i.e., CST-Voting, DTCo and EnSL which utilize simple voting methodologies. The implementation code was written in JAVA, making use of the WEKA 3.9 Machine Learning Toolkit [27].

The performance of the classification algorithms is evaluated using the following performance metrics: F -measure (F_1) and Accuracy (Acc). It is worth mentioning that 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 classification algorithms were evaluated utilizing the chest X-ray (Pneumonia) dataset, the Shenzhen lung mask (Tuberculosis) dataset and the CT Medical images dataset.

- *Chest X-ray (Pneumonia) dataset:* The 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. 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) as in [28].
- *Shenzhen lung mask (Tuberculosis) dataset:* Shenzhen Hospital is one of the largest hospitals in China for infectious diseases with a focus both on their prevention, as well as treatment. The X-rays were collected within a one-month period, mostly in September 2012, as a part of the daily routine, using a Philips DR Digital Diagnost system. The dataset was constructed by manually-segmented lung masks for the Shenzhen Hospital X-ray set as presented in [29]. These segmented lung masks were originally 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" [29]. The set contained 279 normal CXRs and 287 abnormal ones with tuberculosis. All classification algorithms were evaluated using the stratified ten-fold cross-validation.
- *CT Medical images dataset:* This data collection contains 100 images [30] 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* [31]. The images consist of the middle slice of all Computed Tomography (CT) images taken from 69 different patients. The dataset is designed 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 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). All classification algorithms were evaluated using the stratified ten-fold cross-validation.

The training partition was randomly divided into labeled and unlabeled subsets. In order to study the influence of the amount of labeled data, four different ratios (R) of the training data were used: 10%, 20%, 30% and 40%. Using the recommendation established in [9,32] in the division process

we do not maintain the class proportion in the labeled and unlabeled sets since the main aim of semi-supervised classification is to exploit unlabeled data for better classification results. Hence, we use a random selection of examples that will be marked as labeled instances, and the class label of the rest of the instances will be removed. Furthermore, we ensure that every class has at least one representative instance.

4.2. Performance Evaluation of WvEnSL against Ensemble Self-Labeled Algorithms

Next, we focus our interest on the experimental analysis for evaluating the classification performance of WvEnSL algorithm against the ensemble self-labeled algorithms CST-Voting and DTCo, which utilize simple voting methodologies. It is worth noticing that our main goal is to measure the effectiveness of the proposed weighted voting strategy over the simple majority voting; therefore, we will compare ensembles using identical set of classifiers. This will eliminate the source of discrepancy originated from unequal classifiers. Thus, the difference in accuracy can solely be attributed to the difference of voting methodologies.

Furthermore, the base learners utilized in all self-labeled algorithms are the Sequential Minimum Optimization (SMO) [33], the C4.5 decision tree algorithm [34] and the k NN algorithm [35] as in [2,7–9], which probably constitute the most effective and popular machine learning algorithms for classification problems [36].

- “CST-Voting (SMO)” stands for an ensemble of Co-training, Self-training and Tri-training with SMO as base learner using majority voting [2].
- “WvEnSL₁ (SMO)” stands for Algorithm WvEnSL using the same components classifiers as CST-Voting (SMO).
- “CST-Voting (C4.5)” stands for an ensemble of Co-training, Self-training and Tri-training with C4.5 as base learner using majority voting [2].
- “WvEnSL₁ (C4.5)” stands for Algorithm WvEnSL using the same components classifiers as CST-Voting (C4.5).
- “CST-Voting (k NN)” stands for an ensemble of Co-training, Self-training and Tri-training with k NN as base learner using majority voting [2].
- “WvEnSL₁ (k NN)” stands for Algorithm WvEnSL using the same components classifiers as CST-Voting (k NN).
- “DTCo” stands for an ensemble of Democratic-Co learning, Tri-training and Co-Bagging with C4.5 as base learner using majority voting [16].
- “WvEnSL₂” stands for Algorithm WvEnSL using the same components classifiers as DTCo.
- “EnSL” stands for an ensemble of Self-training, Democratic-Co learning, Tri-training and Co-Bagging with C4.5 as base learner using majority voting [17].
- “WvEnSL₃” stands for Algorithm WvEnSL using the same components classifiers as EnSL.

The configuration parameters for all supervised classifiers and self-labeled algorithms, utilized in our experiments, are presented in Table 1.

Table 1. Parameter specification for all the base learners and self-labeled methods used in the experimentation.

Algorithm		Parameters
SMO	Supervised base learner	$C = 1.0$, Tolerance parameter = 0.001, Pearson VII function-based kernel, $\text{Epsilon} = 1.0 \times 10^{-12}$, Fit logistic models = true.
C4.5	Supervised base learner	Confidence level: $c = 0.25$, Minimum number of item-sets per leaf: $i = 2$, Prune after the tree building.
kNN	Supervised base learner	Number of neighbors = 3, Euclidean distance.
Self-training	Self-labeled (single classifier)	MaxIter = 40, $c = 95\%$.
Co-training	Self-labeled (multiple classifier)	MaxIter = 40, Initial unlabeled pool = 75
Tri-training	Self-labeled (multiple classifier)	No parameters specified.
Co-Bagging	Self-labeled (multiple classifier)	Committee members = 3, Ensemble learning = Bagging.
Democratic-Co	Self-labeled (multiple classifier)	Classifiers = kNN, C4.5, NB.
CST-Voting	Ensemble of self-labeled	No parameters specified.
DTCo	Ensemble of self-labeled	No parameters specified.
EnSL	Ensemble of self-labeled	No parameters specified.

Tables 2–4 presents the performance of all ensemble self-labeled methods on Pneumonia dataset, Tuberculosis dataset and CT Medical dataset, respectively. Notice that the highest classification performance for each ensemble of classifiers and performance metric is highlighted in bold. The aggregated results showed that the new weighted voting strategy exploits the individual predictions of each component classifier more efficiently than the simple voting schemes, illustrating better classification performance. WvEnSL₃ exhibits the best performance, reporting the highest F_1 -score and accuracy, relative to all classification benchmarks and labeled ratio, followed by WvEnSL₂. In more detail, WvEnSL₃ demonstrates 82.53–83.49%, 69.79–71.73% and 69–77% classification accuracy for Pneumonia dataset, Tuberculosis dataset and CT Medical dataset, respectively; while WvEnSL₂ reports 81.89–83.17%, 69.79–71.55% and 67–77%, in the same situations.

The statistical comparison of several classification algorithms over multiple datasets is fundamental in the area of machine learning and it is usually performed by means of a statistical test [2,7–9]. 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) [37] test.

Table 2. Performance evaluation of WvEnSL against ensemble self-labeled algorithms for Pneumonia dataset.

Algorithm	Ratio = 10%		Ratio = 20%		Ratio = 30%		Ratio = 40%	
	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc
CST-Voting (SMO)	83.08%	75.32%	83.26%	75.64%	83.39%	75.80%	83.39%	75.80%
WvEnSL ₁ (SMO)	83.39%	75.80%	83.48%	75.96%	83.76%	76.44%	83.85%	76.60%
CST-Voting (C4.5)	85.52%	79.97%	85.85%	80.45%	86.68%	81.73%	86.58%	81.57%
WvEnSL ₁ (C4.5)	85.65%	80.13%	86.08%	80.77%	86.78%	81.89%	86.92%	82.05%
CST-Voting (kNN)	82.91%	75.48%	83.09%	75.80%	83.15%	75.96%	83.73%	76.76%
WvEnSL ₁ (kNN)	83.63%	76.60%	83.73%	76.76%	83.95%	77.08%	84.23%	77.56%
DTCo	86.79%	81.41%	87.21%	82.05%	87.21%	82.05%	87.74%	82.85%
WvEnSL ₂	87.12%	81.89%	87.44%	82.37%	87.54%	82.53%	87.97%	83.17%
EnSL	87.19%	82.05%	86.92%	81.57%	87.34%	82.21%	87.61%	82.69%
WvEnSL ₃	87.51%	82.53%	87.70%	82.69%	88.23%	83.49%	88.17%	83.49%

Table 3. Performance evaluation of WvEnSL against ensemble self-labeled algorithms for Tuberculosis dataset.

Algorithm	Ratio = 10%		Ratio = 20%		Ratio = 30%		Ratio = 40%	
	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc
CST-Voting (SMO)	69.27%	69.43%	68.65%	68.37%	69.50%	69.61%	70.42%	70.32%
WvEnSL ₁ (SMO)	69.73%	69.79%	69.73%	69.79%	70.32%	70.32%	71.00%	70.85%
CST-Voting (C4.5)	66.67%	67.31%	68.19%	68.02%	67.51%	68.20%	69.52%	69.79%
WvEnSL ₁ (C4.5)	67.86%	68.20%	69.26%	69.26%	69.63%	69.79%	69.98%	70.14%
CST-Voting (kNN)	65.71%	66.08%	66.43%	66.96%	68.21%	68.55%	68.93%	69.26%
WvEnSL ₁ (kNN)	65.83%	66.25%	67.14%	67.49%	68.57%	68.90%	69.40%	69.61%
DTCo	69.73%	69.79%	69.96%	69.96%	71.45%	71.20%	71.80%	71.55%
WvEnSL ₂	69.73%	69.79%	70.19%	70.14%	71.58%	71.38%	71.80%	71.55%
EnSL	69.73%	69.79%	69.96%	69.96%	71.00%	70.85%	71.58%	71.38%
WvEnSL ₃	69.73%	69.79%	70.19%	70.14%	71.58%	71.38%	72.03%	71.73%

Table 4. Performance evaluation of WvEnSL against ensemble self-labeled algorithms for CT Medical dataset.

Algorithm	Ratio = 10%		Ratio = 20%		Ratio = 30%		Ratio = 40%	
	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc
CST-Voting (SMO)	66.67%	66.00%	70.00%	70.00%	73.08%	72.00%	75.00%	74.00%
WvEnSL ₁ (SMO)	68.00%	68.00%	71.29%	71.00%	73.79%	73.00%	75.73%	75.00%
CST-Voting (C4.5)	67.96%	67.00%	71.84%	71.00%	73.79%	73.00%	73.79%	73.00%
WvEnSL ₁ (C4.5)	69.90%	69.00%	73.79%	73.00%	75.00%	74.00%	75.73%	75.00%
CST-Voting (kNN)	66.00%	66.00%	69.90%	69.00%	73.79%	73.00%	73.27%	73.00%
WvEnSL ₁ (kNN)	66.67%	67.00%	70.59%	70.00%	72.00%	72.00%	74.75%	75.00%
DTCo	66.02%	65.00%	69.90%	69.00%	72.55%	72.00%	74.29%	73.00%
WvEnSL ₂	67.33%	67.00%	71.29%	71.00%	72.55%	72.00%	76.92%	76.00%
EnSL	64.08%	63.00%	71.84%	71.00%	74.29%	73.00%	74.29%	73.00%
WvEnSL ₃	69.90%	69.00%	75.73%	75.00%	76.47%	76.00%	77.67%	77.00%

Let r_i^j be the rank of the j -th of k learning algorithms on the i -th of M problems. Under the null-hypothesis H_0 , which states that all the algorithms are equivalent, the Friedman aligned ranks test statistic is defined by:

$$F_{AR} = \frac{(k-1) \left[\sum_{j=1}^k \hat{R}_j^2 - (kM^2/4)(kM+1)^2 \right]}{\frac{kM(kM+1)(2kM+1)}{6} - \frac{1}{k} \sum_{i=1}^M \hat{R}_i^2}$$

where \hat{R}_i is equal to the rank total of the i -th dataset and \hat{R}_j is the rank total of the j -th algorithm. The test statistic F_{AR} is compared with the χ^2 distribution with $(k-1)$ degrees of freedom. It is worth noticing that, FAR test does not require the commensurability of the measures across different datasets, since it is non-parametric, neither assumes the normality of the sample means, and thus, it is robust to outliers.

Additionally, in order to identify which algorithms report significant differences, the Finner test [38] with a significance level $\alpha = 0.05$, is applied as a post-hoc procedure. More analytically, the Finner procedure adjusts the value of α in a step-down manner. Let p_1, p_2, \dots, p_{k-1} be the ordered p -values with $p_1 \leq p_2 \leq \dots \leq p_{k-1}$ and H_1, H_2, \dots, H_{k-1} be the corresponding hypothesis. The Finner procedure rejects H_1-H_{i-1} if i is the smallest integer such that $p_i > 1 - (1 - \alpha)^{(k-1)/i}$, while the adjusted Finner p -value is defined by:

$$p_F = \min \left\{ 1, \max \left\{ 1 - (1 - p_j)^{(k-1)/j} \right\} \right\},$$

where p_j is the p -value obtained for the j -th hypothesis and $1 \leq j \leq i$. It is worth mentioning that the test rejects the hypothesis of equality when the p_F is less than α .

The control algorithm for the post-hoc test is determined by the best (lowest) ranking obtained in each FAR test. Moreover, the adjusted p -value with Finner's test (p_F) was presented based on the corresponding control algorithm at the α level of significance while the post-hoc test rejects the hypothesis of equality when the value of p_F is less than the value of α . It is worth mentioning that the FAR test and the Finner post-hoc test were performed based on the classification accuracy of each algorithm over all datasets and labeled ratio.

Table 5 presents the information of the statistical analysis performed by nonparametric multiple comparison procedures for all ensemble self-labeled algorithms. The interpretation of Table 5 demonstrates that WvEnSL₃ reports the highest probability-based ranking by statistically presenting better results, followed by WvEnSL₂ and WvEnSL₁ (C4.5). Moreover, it is worth mentioning that all weighted voting ensemble outperformed the corresponding ensemble which utilize classical voting schemes. Finally, based on the statistical analysis, we can easily conclude that the new weighted voting scheme had a significant impact on the performance of all ensemble of self-labeled algorithms.

Table 5. Friedman Aligned Ranking (FAR) test and Finner post-hoc test.

Algorithm	FAR	Finner Post-Hoc Test	
		p_F -Value	Null Hypothesis
WvEnSL ₃	15.667	-	-
WvEnSL ₂	34.958	0.174312	accepted
WvEnSL ₁ (C4.5)	44.208	0.049863	rejected
EnSL	47.958	0.029437	rejected
DTC _o	51.125	0.018734	rejected
CST-Voting (C4.5)	64.042	0.001184	rejected
WvEnSL ₁ (SMO)	71.417	0.000194	rejected
CST-Voting (SMO)	88.292	0.000001	rejected
WvEnSL ₁ (kNN)	89.083	0.000001	rejected
CST-Voting (kNN)	98.250	0.000001	rejected

4.3. Performance Evaluation of WvEnSL against Classical Supervised Algorithms

Next, we compare the classification performance of the proposed algorithm against the classical supervised classification algorithms: SMO, C4.5 and *k*NN. Moreover, we compare the performance of iCST-Voting against the ensemble of classifiers (Voting) which combines the individual predictions of the supervised classifiers utilizing a simple majority voting strategy. It is worth noticing that

- we selected WvEnSL₃ from all versions of the proposed algorithm since it presented the best overall performance.
- all supervised algorithms were trained using with 100% of the training set while WvEnSL₃ was trained using $R = 40\%$ of the training set.

Table 6 presents the performance of the proposed algorithm WvEnSL₃ against the supervised algorithms SMO, C4.5, *k*NN and Voting on Pneumonia dataset, Tuberculosis dataset and CT Medical dataset. As above mentioned, the highest classification performance for each labeled ratio and performance metric is highlighted in bold. The aggregated results show that WvEnSL₃ is the most efficient algorithm since it illustrates the best overall classification performance. More specifically, WvEnSL₃ exhibits the highest F_1 -score and classification accuracy on Pneumonia and Tuberculosis datasets, while for CT Medical dataset, WvEnSL₃ reports the second best performance, considerably outperformed by C4.5.

Table 6. Performance evaluation WvEnSL₃ against state-of-the-art supervised algorithms on Pneumonia dataset, Tuberculosis dataset and CT Medical dataset.

Algorithm	Pneumonia		Tuberculosis		CT Medical	
	F_1	Acc	F_1	Acc	F_1	Acc
SMO	74.03%	76.76%	71.41%	71.37%	74.91%	75.00%
C4.5	72.41%	74.83%	62.32%	62.36%	79.82%	80.00%
3NN	72.32%	74.51%	67.51%	67.49%	67.08%	67.00%
Voting	73.34%	76.12%	71.00%	71.02%	74.07%	74.00%
WvEnSL ₃	88.17%	83.49%	72.03%	71.73%	77.67%	77.00%

5. Conclusions

In this work, we proposed a new weighted voting ensemble self-labeled algorithm for the detection of lung abnormalities from X-rays, entitled WvEnSL. The proposed algorithm combines the individual predictions of self-labeled algorithms utilizing a new weighted voting methodology. The significant advantage of WvEnSL is that weights assigned on each component classifier of the ensemble are based on its accuracy on each class of the dataset.

For testing purposes, the algorithm was extensively evaluated using the chest X-rays (Pneumonia) dataset, the Shenzhen lung mask (Tuberculosis) dataset and the CT Medical images dataset. Our

numerical experiments indicated better classification accuracy of the WvEnSL and demonstrated the efficiency of the new weighted voting scheme, as statistically confirmed by the Friedman Aligned Ranks nonparametric test as well as the Finner post hoc test. Therefore, we can conclude that the new weighted voting strategy had a significant impact on the performance of all ensembles of self-labeled algorithms, exploiting the individual predictions of each component classifier more efficiently than the simple voting schemes. Finally, it is worth mentioning that efficient and powerful classification models could be developed by the adaptation of ensemble methodologies in the SSL framework.

In our future work, we intend to pursue extensive empirical experiments to compare the proposed WvEnSL with other algorithms belonging to different SSL classes, and evaluate its performance using various component self-labeled algorithms and base learners. Furthermore, since our preliminary numerical experiments are quite encouraging, our next step is to explore the performance of the proposed algorithm on imbalanced datasets [39,40] and incorporate our proposed methodology for multi-target problems [41–43]. Additionally, another interesting aspect is the use of other component classifiers in the ensemble and enhance our proposed framework with more sophisticated and theoretically sound criteria for the development of an advanced weighted voting strategy. Finally, we intend to investigate and evaluate different strategies for the selection of the evaluation set.

Author Contributions: I.E.L., A.K., V.T. and P.P. conceived of the idea, designed and performed the experiments, analyzed the results, drafted the initial manuscript and revised the final manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Van Ginneken, B.; Stegmann, M.B.; Loog, M. Segmentation of anatomical structures in chest radiographs using supervised methods: A comparative study on a public database. *Medical Image Anal.* **2006**, *10*, 19–40. [[CrossRef](#)] [[PubMed](#)]
2. Livieris, I.; Kanavos, A.; Tampakas, V.; Pintelas, P. An ensemble SSL algorithm for efficient chest X-ray image classification. *J. Imaging* **2018**, *4*, 95. [[CrossRef](#)]
3. Zhu, X.; Goldberg, A. Introduction to semi-supervised learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **2009**, *3*, 1–130. [[CrossRef](#)]
4. Chapelle, O.; Scholkopf, B.; Zien, A. Semi-supervised learning. *IEEE Trans. Neural Netw.* **2009**, *20*, 542–542. [[CrossRef](#)]
5. Levatic, J.; Dzeroski, S.; Supek, F.; Smuc, T. Semi-supervised learning for quantitative structure-activity modeling. *Informatica* **2013**, *37*, 173–179.
6. Levatić, J.; Ceci, M.; Kocev, D.; Džeroski, S. Semi-supervised classification trees. *J. Intell. Inf. Syst.* **2017**, *49*, 461–486. [[CrossRef](#)]
7. Livieris, I.; Kanavos, A.; Tampakas, V.; Pintelas, P. An auto-adjustable semi-supervised self-training algorithm. *Algorithm* **2018**, *11*, 139. [[CrossRef](#)]
8. Livieris, I.; Kiriakidou, N.; Kanavos, A.; Tampakas, V.; Pintelas, P. On ensemble SSL algorithms for credit scoring problem. *Informatics* **2018**, *5*, 40. [[CrossRef](#)]
9. Triguero, I.; García, S.; Herrera, F. Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study. *Knowl. Inf. Syst.* **2015**, *42*, 245–284. [[CrossRef](#)]
10. Yarowsky, D. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Annual Meeting of the Association For Computational Linguistics, Cambridge, MA, USA, 26–30 June 1995; pp. 189–196.
11. Blum, A.; Mitchell, T. Combining labeled and unlabeled data with co-training. In Proceedings of the 11th Annual Conference on Computational Learning Theory, Madison, WI, USA, 24–26 July 1998; pp. 92–100.
12. Zhou, Y.; Goldman, S. Democratic co-learning. In Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Boca Raton, FL, USA, 15–17 November 2004; IEEE: Piscataway, NJ, USA, 2004; pp. 594–602.

13. Zhou, Z.; Li, M. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 1529–1541. [[CrossRef](#)]
14. Li, M.; Zhou, Z. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **2007**, *37*, 1088–1098. [[CrossRef](#)]
15. Hady, M.; Schwenker, F. Combining committee-based semi-supervised learning and active learning. *J. Comput. Sci. Technol.* **2010**, *25*, 681–698. [[CrossRef](#)]
16. Livieris, I.; Kotsilieris, T.; Anagnostopoulos, I.; Tampakas, V. DTCO: An ensemble SSL algorithm for X-rays classification. In *Advances in Experimental Medicine and Biology*; Springer: Berlin/Heidelberg, Germany, 2018.
17. Livieris, I.; Kanavos, A.; Pintelas, P. Detecting lung abnormalities from X-rays using and improved SSL algorithm. *Electron. Notes Theor. Comput. Sci.* **2019**, accepted for publication.
18. Livieris, I. A new ensemble self-labeled semi-supervised algorithm. *Informatica* **2018**, accepted for publication.
19. Jaeger, S.; Karargyris, A.; Candemir, S.; Folio, L.; Siegelman, J.; Callaghan, F.; Xue, Z.; Palaniappan, K.; Singh, R.; Antani, S.; et al. Automatic tuberculosis screening using chest radiographs. *IEEE Trans. Med. Imaging* **2014**, *33*, 233–245. [[CrossRef](#)]
20. Melendez, J.; van Ginneken, B.; Maduskar, P.; Philipsen, R.; Reither, K.; Breuninger, M.; Adetifa, I.; Maane, R.; Ayles, H.; Sánchez, C. A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest X-rays. *IEEE Trans. Med. Imaging* **2015**, *34*, 179–192. [[CrossRef](#)]
21. Alam, J.; Alam, S.; Hossain, A. Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifier. In Proceedings of the 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering, Rajshahi, Bangladesh, 8–9 February 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–4.
22. Madani, A.; Moradi, M.; Karargyris, A.; Syeda-Mahmood, T. Semi-supervised learning with generative adversarial networks for chest X-ray classification with ability of data domain adaptation. In Proceedings of the 15th IEEE International Symposium on Biomedical Imaging, Washington, DC, USA, 4–7 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1038–1042.
23. Guan, Q.; Huang, Y. Multi-label chest X-ray image classification via category-wise residual attention learning. *Pattern Recognit. Lett.* **2018**, doi:10.1016/j.patrec.2018.10.027. [[CrossRef](#)]
24. Dietterich, T. Ensemble methods in machine learning. In *Multiple Classifier Systems*; Kittler, J., Roli, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2001; Volume 1857, pp. 1–15.
25. Rokach, L. *Pattern Classification Using Ensemble Methods*; World Scientific Publishing Company: Singapore, 2010.
26. Powers, D.M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *J. Mach. Learn. Technol.* **2011**, *2*, 37–63.
27. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. The WEKA data mining software: An update. *SIGKDD Explor. Newsl.* **2009**, *11*, 10–18. [[CrossRef](#)]
28. Kermany, D.; Goldbaum, M.; Cai, W.; Valentim, C.; Liang, H.; Baxter, S.; McKeown, A.; Yang, G.; Wu, X.; Yan, F. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **2018**, *172*, 1122–1131. [[CrossRef](#)]
29. Stirenko, S.; Kochura, Y.; Alienin, O.; Rokovyi, O.; Gang, P.; Zeng, W.; Gordienko, Y. Chest X-ray analysis of tuberculosis by deep learning with segmentation and augmentation. *arXiv* **2018**, arXiv:1803.01199.
30. Albertina, B.; Watson, M.; Holback, C.; Jarosz, R.; Kirk, S.; Lee, Y.; Lemmerman, J. Radiology data from the cancer Genome Atlas Lung Adenocarcinoma [TCGA-LUAD] collection. *Cancer Imaging Arch.* **2016**.
31. Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **2013**, *26*, 1045–1057. [[CrossRef](#)]
32. Wang, Y.; Xu, X.; Zhao, H.; Hua, Z. Semi-supervised learning based on nearest neighbor rule and cut edges. *Knowl.-Based Syst.* **2010**, *23*, 547–554. [[CrossRef](#)]
33. Platt, J. *Advances in Kernel Methods—Support Vector Learning*; MIT Press: Cambridge, MA, USA, 1998.
34. Quinlan, J. *C4.5: Programs for Machine Learning*; Morgan Kaufmann: San Francisco, CA, USA, 1993.
35. Aha, D. *Lazy Learning*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1997.
36. Wu, X.; Kumar, V.; Quinlan, J.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.; Ng, A.; Liu, B.; Yu, P.; et al. Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **2008**, *14*, 1–37. [[CrossRef](#)]

37. Hodges, J.; Lehmann, E. Rank methods for combination of independent experiments in analysis of variance. *Ann. Math. Stat.* **1962**, *33*, 482–497. [[CrossRef](#)]
38. Finner, H. On a monotonicity problem in step-down multiple test procedures. *J. Am. Stat. Assoc.* **1993**, *88*, 920–923. [[CrossRef](#)]
39. Li, S.; Wang, Z.; Zhou, G.; Lee, S. Semi-supervised learning for imbalanced sentiment classification. In Proceedings of the IJCAI Proceedings-International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, 16–22 July 2011; Volume 22, p. 1826.
40. Jeni, L.A.; Cohn, J.F.; De La Torre, F. Facing imbalanced data—Recommendations for the use of performance metrics. In Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland, 2–5 September 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 245–251.
41. Levatić, J.; Ceci, M.; Kocev, D.; Džeroski, S. Self-training for multi-target regression with tree ensembles. *Knowl.-Based Syst.* **2017**, *123*, 41–60. [[CrossRef](#)]
42. Levatić, J.; Kocev, D.; Džeroski, S. The importance of the label hierarchy in hierarchical multi-label classification. *J. Intell. Inf. Syst.* **2015**, *45*, 247–271. [[CrossRef](#)]
43. Levatić, J.; Kocev, D.; Ceci, M.; Džeroski, S. Semi-supervised trees for multi-target regression. *Inf. Sci.* **2018**, *450*, 109–127. [[CrossRef](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).