Learning and Nonlinear Models - Journal of the Brazilian Society on Computational Intelligence (SBIC), Vol. 20, Iss. 2, pp. 63-73, 2022

APPLYING THE LIFELONG MACHINE LEARNING PARADIGM IN TUBERCULOSIS TRIAGE

Regina Reis da Costa Alves
Programa de Engenharia Biomedica (COPPE/UFRJ)
regina.alves@poli.ufrj.br

Frederico Caetano Jandre de Assis Tavares, Anete Trajman, Jose Manoel de Seixas
Programa de Engenharia Biomedica (COPPE), Faculdade de Medicina (UFRJ), Laboratorio de Processamento de Sinais (COPPE)
jandre@peb.ufrj.br, atrajman@gmail.com, seixas@lps.ufrj.br

Abstract – Tuberculosis (TB) and pneumonia, including pneumonia from SARS-CoV-2 infection, are among the main causes of lower respiratory infections, which are the fourth cause of death worldwide. Recently, the World Health Organization recommended the use of computer-aided diagnosis (CAD) software as a tool to analyze chest radiographs (CXR) for TB screening and triage. Most CAD developed to date aim to screen exclusively for TB. This work applies the lifelong machine learning paradigm to detect both pneumonia and TB through CXRs and evaluate the models' ability to retain and acquire knowledge. Two well-known lifelong learning models, the Efficient Lifelong Learning Algorithm (ELLA) and Learning without Forgetting (LwF), were applied to two public CXR datasets containing TB and pneumonia samples together with healthy CXR samples. Pneumonia detection was learned first and TB detection was learned as second task. The SP index, a function of sensitivity and specificity, was used to evaluate the models. We concluded that both algorithms were able to retain knowledge about pneumonia detection and were also able to learn TB detection.

Keywords – lifelong machine learning, continuous learning, tuberculosis, chest radiographs, pneumonia

1. INTRODUCTION

Lower respiratory infections are the fourth cause of death worldwide [1]. Tuberculosis (TB) and bacterial or viral pneumonia, including pneumonia from SARS-CoV-2 infection, are among the main causes of lower respiratory infections [2]. TB is caused by the bacillus *Mycobacterium tuberculosis* and in general affects the lungs, but can also affect other organs [3]. It has killed 1.3 million people in 2020. For the first time in more than 10 years, this number have increased, as the COVID-19 pandemic caused a reduction in TB diagnosis and treatment [3]. TB is endemic and in the long term, it will probably continue to kill more than COVID-19.

Chest radiographs (CXRs) are recommended by the World Health Organization (WHO) as a triage tool for TB [4]. Nevertheless, a shortage of skilled radiologists in many high TB-burden countries and the discordance of human CXR readers on TB diagnosis limit its use [5]. Thus, in March 2021, the WHO included in the TB Screening Guidelines the use of computer-aided diagnosis (CAD) software as an alternative to analyzing digital CXR for TB screening and triage in individuals aged 15 years old and above [6].

Most available CADs are designed to screen TB only [7], [8], [9]. However, for a CAD system to have clinical utility, it should be able to account for most thoracic abnormalities observed on CXRs, as there are various diseases that can cause the same symptoms as pulmonary TB in real-world clinical practice [10]. Developing models that can detect different diseases is useful, not only for generalizing aims, but also to avoid false positives in TB detection because other diseases may have similar radiological signs [11]. Some examples are different forms of pneumonia, neoplasms, edema and hemorrhagic disease [9]. Also, in clinical practice, patients presenting with respiratory symptoms usually undergo CXR for differential diagnosis of several diseases, not one or another specific disease.

Pneumonia is one of the most frequent thoracic diseases that can mimic TB in CXRs [10]. It is an acute respiratory infection that affects the lungs and can be caused by viruses, bacteria or fungi. The patient's alveoli are filled with pus and fluid. It is the single most frequent infectious cause of pediatric death worldwide, accounting for 14% of all deaths of children under 5 years of age in 2019 [12].

In this work, we propose the application of two algorithms, Efficient Lifelong Learning Algorithm (ELLA) [13] and Learning without Forgetting (LwF) [14], both developed in the lifelong machine learning paradigm (LML) [15], to detect pneumonia and TB through CXRs. LML corresponds to systems that can learn many tasks over time while retaining the knowledge from old tasks [15].This would represent an advance vis-a-vis the existing CAD models that screen TB only and the models that screen more diseases but do not learn new diseases over time.

2. TUBERCULOSIS AND MACHINE LEARNING

The radiological signs of pulmonary TB are diverse, which adds complexity to TB diagnosis [9]. In CXR, TB disease is characterized by the presence of consolidations, fibrosis, calcification and cavitary lesions in the lungs [16]. Pleural effusion is also a common radiological sign of intrathoracic TB [17].

Several artificial intelligence-based computer programs were developed aiming to analyze CXRs for pulmonary tuberculosis. The AI4HLTH resource centre from the Stop TB Partnership and FIND keeps a list with all CAD products that may be used for TB detection in all stages of development [18]. The list contains information about accuracy, operational characteristics, options for integration into the legacy system, costs, data sharing and privacy aspects, among others [5].

There are conflicting data on these CAD's accuracy. In an external evaluation of five commercial CADs for triaging TB, all outperformed experienced Bangladeshi human readers for detecting TB in CXRs [19]. Two CAD products - qXR and CAD4TB - met WHO-recommended minimal accuracy for TB triage, i.e., 90% sensitivity and 70% specificity [20], when applied to symptomatic adults at the Indus Hospital in Pakistan. A lower sensitivity was observed for smear-negative pulmonary TB. However, when applied to clinical data from 6 source studies, 64% of whom with microbiological test information, these two CADs did not met the targets established by the WHO for a triage test [21]. Lunit Insight was also below WHO targets.

The deep convolutional neural network (CNN) topologies have become the preferred technique for general-purpose image processing, including medical images [22]. A CNN based on the architecture of *AlexNet* [23], trained with datasets from Shenzhen, Montgomery [24] and the Korean Institute of Tuberculosis (KIT), achieved an average accuracy of 90.3% for TB detection [25]. Two other CNNs, *AlexNet* and *GoogLeNet*, added the classification by a radiologist to decide on disagreement between the two CADS, which improved the sensitivity to 97.3% and specificity of 100% [26]. They were trained with the Montgomery, Shenzhen, Belarus and the Thomas Jefferson University Hospital datasets.

If only TB is considered in diagnosis, other diseases may be misclassified as TB. A model trained with a TB-specific dataset and tested with the ChestX-ray8 dataset, which has 8 classes [11], classified 36.51% of abnormal radiographs in the ChestX-ray8 dataset as TB. The authors concluded that there was an over-diagnosis of TB. Although many CXR abnormalities are common in TB and other lung diseases, few models have been developed with the aim of diagnosing TB and other diseases. An accuracy of 98.3% was obtained with a model for TB, pneumonia, malignant lung neoplasms and pneumothorax detection, based on CNNs with classifiers in parallel [10]. A downside of this approach is that the whole diseases' datasets must be available at the beginning of the training, and if a new disease is presented, the model has to be retrained from the start. This is not the case if the LML paradigm is applied.

3. THE LIFELONG MACHINE LEARNING PARADIGM

LML corresponds to systems that can learn many tasks over time from one or more domains, while retaining the knowledge from old tasks, that is, avoiding catastrophic forgetting, and using it to learn new tasks more efficiently and effectively [15]. It is also known as continuous learning and continual learning. Systems should preferably present forward transfer (better performance and faster learning on new tasks) and backward transfer (better performance on previous tasks) because of shared knowledge between tasks; it is also desirable that systems have minimal access to previous tasks and a minimal increase in model capacity and computation when learning new tasks [27].

In practice, it is a challenge to have, at the same time, a system that does not forget old tasks and learns fast and effectively new tasks. This capability is present in the human brain. The process in which the brain adapts to environmental changes is called plasticity and was first demonstrated by the neuroanatomist Michele Vincenzo Malacarne in 1783 [28]. At the same time, the brain needs to be stable enough to retain knowledge over time. This trade-off is known as the plasticity-stability dilemma [29]. The same dilemma is present in neural networks, as a network that is too plastic will tend to overwrite its weights when new tasks are learned and deteriorate performance in old tasks, and a very stable network will change very little its weights and tend to perform poorly in new tasks.

The application of this paradigm in medicine and, specifically in CXR analysis, is still limited. An application was made in a simulated scenario in which data are spread across multiple hospitals and a machine learning model must be trained sequentially [30]. The model was tested on the CXR dataset for Tuberculosis at the Korean Tuberculosis Institute [25]. Another study involved the diagnosis of congenital heart disease in fetuses, using small regions of interest in medical images [31].

There are different ways to apply the LML paradigm. The strategies of continuous learning algorithms can be grouped into three categories: architectural, regularization and rehearsal strategies [32], detailed below.

3.1 Architectural strategy

The strategy is based on models that use specific architectures, addition of layers and parameters, weight-freezing strategies, among others, in order to avoid catastrophic forgetting. The Progressive Neural Network (PNN) algorithm [33] is an example of this strategy. In this algorithm, the network is initialized with a single column, which consists of a deep neural network with L layers. When a second task arrives, a new column is added and the model learns lateral connections between the two columns, in order to extract useful characteristics for the new task.

Another example is an algorithm developed for building dynamically expanding networks [34]. When a new task arrives, firstly the network makes maximum use of the knowledge of previous tasks, retraining a subset of parameters in order to learn the new task. It is then expanded when the accumulated knowledge alone cannot perform satisfactorily on the new task. Unlike

PNN, which always adds a fixed number of neurons, in this case, the number of neurons to be added is dynamically decided for each layer and task.

3.2 Regularization strategy

In this strategy, the loss function encompasses terms that lead to selective consolidation of weights that are important to retain prior knowledge. The algorithm called *Elastic Weight Consolidation* (EWC) [35] is an example, which imposes that important parameters of the neural network remain close to their original values. This is done by including a penalty in the objective function, calculated by Fisher's information matrix, which is a measure of the amount of information that the viewed data bring from a parameter. The more information there is about the parameter, the greater the penalty associated with changing it.

The algorithm called AR1, for multiclass problems, combines the architectural and regularization strategy of penalizing the objective function and the growth of the network [32]. It is based on a deep network that has a parameter set shared among all classes. For each batch of training, the output layer is expanded with neurons corresponding to the number of new classes in this batch. The shared weights are also updated, but variations on these weights are penalized in the objective function. The magnitude of the penalty varies according to the importance of the weight in the output of the previous tasks, using the Synaptic Intelligence technique [36].

3.3 Rehearsal strategy

This strategy is based on re-presenting periodically previous information to the model, in order to strengthen the memory of what has already been learned, either by storing part of the data referring to previous tasks or by using generative models to generate pseudo-data. The iCaRL algorithm [37] is an example of this strategy, applied to the incremental learning of classes. It is based on a CNN, which is trained to extract features from elements. To predict the class of an input, a prototype vector is computed for each class already observed. These vectors are compared to the feature vector related to the image that is to be classified, which the network also computes, and the class defined for the image is the one whose prototype vector is most similar. When a new class is presented to the network, first the outputs for the classes which were already learned are stored. Then, the CNN parameters are updated by minimizing a loss function with the restriction that the results for the old classes are reproduced.

Another algorithm uses an adversarial deep generating network (GAN) [38], which is trained to generate data similar to those referring to past tasks [39]. When a new task is presented, the GAN generates as much data as necessary and merges it with the new data to update both the GAN and the network that acts as a solver.

4 DESCRIPTION OF THE APPLIED ALGORITHMS

The first LML algorithm applied is called Efficient Lifelong Learning Algorithm (ELLA) [13], which uses the architectural strategy. This algorithm was chosen beacuse it is one of the important pioneers in the LML paradigm and because it is based on simple models, either linear regression or logistic regression. In the present application, as it involves classification, the algorithm is based on logistic regression.

ELLA assumes that the models' parameters can be represented as a linear combination of shared latent components from a knowledge repository. The library of latent components is the main mechanism of knowledge retention and transference, as it is shared between the specific models associated with each task. The code used in this study was adapted from the code made publicly available by the authors.

Let $L \subseteq \mathbb{R}^{d \times nc}$ be the library of latent components, d the dimension of the inputs x and nc the number of latent components in the library. The vector $\theta^{(t)}$, corresponding to the parameters of each task t, is calculated as a linear combination of the components in L whose weights are given by the vector $s^{(t)} \in \mathbb{R}^{nc}$. That is, for each task t we have that:

$$\theta^{(t)} = Ls^{(t)} \tag{1}$$

Equation 2 shows the objective function to be minimized, depending on L and the $s^{(t)}$ corresponding to each task:

$$\min_{L,s^{(t)}} \frac{1}{N} \sum_{t=1}^{N} \{ \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L}(f(x_i^{(t)}; \theta^{(t)}), y_i^{(t)}) + \mu ||s^{(t)}||_1 \} + \lambda ||L||_F^2 \tag{2}$$

where:

$$f(x_i^{(t)}; \theta^{(t)}) = \frac{1}{1 + e^{-(\theta^{(t)})^T x_i^{(t)}}} \tag{3}$$

and:
N is the number of tasks
n_t is the number of elements corresponding to task t
$x_i^{(t)}, y_i^{(t)}$ is the ith labeled training instance for task t
$\mathcal{L}$ is a known loss function

$||L||_F$ is the Frobenius norm of L

Catastrophic forgetting is avoided by the terms in the objective function which penalize classification errors of past tasks. A disadvantage of this definition is that all of the previous training data are required to calculate the objective function. In order to remove this dependence, the authors use the second-order Taylor expansion around $\hat{\theta}^{(t)} = \arg\min_\theta \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L}(f(x_i^{(t)}; \theta), y_i^{(t)})$, which is an optimal predictor obtained when the model is trained only with training data for task t, leading to the objective function expressed in Equation 4.

$$\min_{L, s^{(t)}} \frac{1}{N} \sum_{t=1}^{N} |||\hat{\theta}^{(t)} - Ls^{(t)}||_{H^{(t)}} + \mu||s^{(t)}||_1 + \lambda||L||_F^2 \tag{4}$$

where:

$$H^{(t)} = \frac{1}{2} \nabla^2_{\theta^{(t)}, \theta^{(t)}} \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L}(f(x_i^{(t)}; \theta^{(t)}), y_i^{(t)})|_{\theta^{(t)} = \hat{\theta}^{(t)}} \tag{5}$$

The objective function is a function of $s^{(t)}$ and L. As the function is not convex in $s^{(t)}$ and L simultaneously, the optimization is done in two steps: first fixing L and optimizing $s^{(t)}$ and then fixing $s^{(t)}$ and optimizing L, as shown in Equations 6 and 7. In order to simplify the optimization process, the $s^{(t)}$'s for the old tasks are kept fixed and only the $s^{(t)}$ for the new task is updated.

$$s^{(t)} \leftarrow \arg\min_{s^{(t)}} ||\hat{\theta}^{(t)} - L_m s^{(t)}||^2_{H^{(t)}} + \mu||s^{(t)}||_1 \tag{6}$$

$$L_{m+1} \leftarrow \arg\min_{L} \frac{1}{N} \sum_{t=1}^{N} (||\hat{\theta}^{(t)} - Ls^{(t)}||^2_{H^{(t)}} + \mu||s^{(t)}||_1) + \lambda||L||_F^2 \tag{7}$$

The fact that the base L changes over time makes possible an improvement in past tasks performance, as their parameters are a function of the base L. As all the tasks' parameters are a linear combination of the base L, the inputs from all tasks must come from the same space $\mathbb{R}^d$.

The second algorithm applied in this study was the *Learning without Forgetting* (LwF) [14], based on a CNN, which is recommended in image classification tasks. It uses the regularization strategy to avoid catastrophic forgetting. The loss function is modified to control forgetting without the need to store the older tasks' datasets.

The network layers are shared by all tasks, except for the last layer. The shared layers' parameters are represented by θ_s and the last layer's parameters are represented by θ_o for the old tasks and θ_n for the new tasks. When a new task arrives, the labeled dataset x_n, y_n associated with it is presented to the network, at a first moment without changing the network's parameters, and the outputs y_o of all the previous tasks' last layer are recorded. Next, the parameters θ_n are added to the output layer, connected to all neurons in the previous layer and initialized with random values.

As the tasks share all the layers except for the last one, all the inputs must come from the same space $\mathbb{R}^d$. The number of classes in each task can vary, as the last layer of the network is task-specific.

The LwF modified loss function is defined as:

$$\min_{\hat{\theta}_s, \hat{\theta}_o, \hat{\theta}_n} \lambda_o \mathcal{L}_{old}(y_o, \hat{y}_o) + \mathcal{L}_{new}(y_n, \hat{y}_n) + \mathcal{R}(\hat{\theta}_s, \hat{\theta}_o, \hat{\theta}_n) \tag{8}$$

where:

$$\mathcal{L}_{old}(y_o, \hat{y}_o) = -H(\hat{y}'_o, \hat{y}'_o) = -\sum_{i=1}^{l} y_o'^{(i)} log \hat{y}_o'^{(i)} \tag{9}$$

$$\mathcal{L}_{new}(y_n, \hat{y}_n) = -y_n \cdot log \hat{y}_n \tag{10}$$

$$y_o'^{(i)} = \frac{(y_o^{(i)})^{1/T}}{\sum_{j=1}^{l} (y_o^{(j)})^{1/T}} \tag{11}$$

$$\hat{y}_o'^{(i)} = \frac{(\hat{y}_o^{(i)})^{1/T}}{\sum_{j=1}^{l} (\hat{y}_o^{(j)})^{1/T}} \tag{12}$$

and:

$\mathcal{R}(\hat{\theta}_s, \hat{\theta}_o, \hat{\theta}_n)$ is a regularization term;

λ_o is a loss balance weight; the higher the value, the more importance is given to the performance in previous tasks, to the detriment of the new task;

$\hat{y}_n$ is the last layer output corresponding to the new task, $\hat{y}_n \equiv CNN(x_n, \hat{\theta}_s, \hat{\theta}_n)$;

y_n is the vector of labels for the new task;

$\hat{y}_o^{(i)}$ is the current last layer output corresponding to the previous task i, $\hat{y}_o^{(i)} \equiv CNN(x_n, \hat{\theta}_s, \hat{\theta}_o^{(i)})$;

$y_o^{(i)}$ is the recorded last layer output corresponding to the previous task i, $y_o^{(i)} \equiv CNN(x_n, \theta_s, \theta_o^{(i)})$;
l is the number of labels;
$y_o'^{(i)}$ and $\hat{y}_o'^{(i)}$ are modified versions of the current and recorded output;
T is a parameter that controls the weight given to output values in the modified version.

5 APPLICATION IN THE TASKS OF DETECTING PNEUMONIA AND TB

The ELLA and LwF algorithms were applied to the classification of two diseases: pneumonia and tuberculosis. Radiological signs associated to TB were already enumerated in section 2. For pneumonia, common signs are lower lobes with bilateral multicentric opacities, diffuse multifocal involvement and pleural effusion [40].

For this purpose, two public CXR datasets were used. The first one is the CheXpert [41], which is a dataset consisting of 224,316 chest radiographs of 65,240 patients who underwent a radiographic examination from Stanford University Medical Center. This dataset presents 14 classes. We selected a subset of images according to the following criteria:

- Only 2 classes: 'Normal' and 'Pneumonia'.
- Images present in the reduced set that was available for download (11 GB, while the full set has 439 GB).
- Front images and AP view.
- Similar amount of images in both classes.

After this process, there were 1,070 images from class 'Normal' and 1,065 images from class 'Pneumonia'.

The second dataset was made available by the U.S. National Library of Medicine and contains radiographs collected at Shenzhen No.3 People's Hospital (China) [24]. It has a total of 662 frontal CXR images, of which 326 correspond to healthy patients and 336 present manifestations of TB. The images are 8-bits RGB with width and height from 948 to 3001 pixels. All the images from this dataset were used in the present study.

A limitation concerning the public datasets used in this work is that they lack information about how the CXRs' diagnosis were made and whether they were confirmed by microbiological tests or by following the patient's clinical outcome.

As the main purpose was to evaluate the application of the algorithms for TB screening and triage, pneumonia detection was defined as the first task to be learned and TB detection as the second task. The reason is that, even with mechanisms to avoid catastrophic forgetting, a decrease in classification performance when learning a new task may occur. Another reason is that the models may benefit from the knowledge acquired when learning to classify pneumonia to improve performance when learning to classify TB [27].

Images for both algorithms were pre-processed, so that all image resolution became 224 x 224 pixels from 3 channels. As the ELLA algorithm is based on logistic regression, the Principal Component Analysis [42] was used in addition to the described pre-processing, in order to reduce dimensionality.

For ELLA, a grid search was made to choose the model hyperparameters with $nc = [2, 4, 8, 12]$, $\mu = [0.1, 1.0, 1.5]$ and $\lambda = [10^{-4}, 10^{-5}, 10^{-6}]$. Different percentages of variance retention from data projection in the principal components were tested (85%, 90% and 95%).

For LwF, the code made available by Mallya *et al* [43] was adapted and a grid search was made for the following hyperparameters, which are related to the CNN training process using Stochastic Gradient Descent (SGD) [44]: learning rate = $[10^{-2}, 10^{-3}, 10^{-4}]$, learning rate decay factor = $[0.05, 0.1, 0.2]$ and weight decay = $[0.00025, 0.0005, 0.001]$.

Three different indexes were chosen as performance indexes: sensitivity and specificity, which are well-known indexes that measure the model performance in detecting, respectively, those with the disease (true positive rate) and those without the disease (true negative rate), and the SP index [45], which is a function of the two former indexes, as shown in Equation 13. This index was chosen because it allows a ballance between sensitivity and specificity and the performance is given by a single value.

$$SP = \sqrt{\sqrt{sensitivity * specificity} * \frac{sensitivity + specificity}{2}} \tag{13}$$

For both algorithms, the cross-validation method with stratified k-folds technique [46] was used in the gridsearch, with k=10. This method gives an estimate of the uncertainty in the results due to the fact that we have a limited sample. The hyperparameters that gave the highest SP index average in the validation sets were chosen. The grid search steps are described in Algorithm 1.

With this procedure, the hyperparameters for ELLA were set as $k = 8$, $\mu = 0.1$ and $\lambda = 10^{-4}$ and for LwF, the learning rate was set as 10^{-3}, the learning rate decay factor as 0.2 and the weight decay as 0.00025.

Then, the ELLA and LwF algorithms with the chosen hyperparameters were applied to the tasks of pneumonia and TB detection. Models were adjusted and results were calculated using Algorithm 2 for LwF, which is similar to Algorithm 1 but presents two main differences: 1) the first "for" loop, which concerned the hyperparameters, was skipped and 2) for each test fold, the validation model with highest SP index was chosen in order to apply it to this test fold. For ELLA, the algorithm was very similar, but there was no need to define a validation set as ELLA is not an iterative algorithm and the hyperparameters have been chosen in the precedent step. The 95% confidence interval (CI) for the three indexes was calculated using the cross-validation method with k-folds technique with k=10.

Algorithm 1 Grid search to choose hyperparameters for ELLA and LwF algorithms

```
for each set of hyperparameters do
    for each task do
        Divide dataset from task in 10 folds
        for each ifold do
            Separate ifold as a test fold
            for each jfold different from ifold do
                Separate jfold as a validation fold
                Train ELLA or LwF model with the other 8 folds
                Apply model to jfold and save in validationResults
                Apply model to the whole dataset and save in allResults
            end for
            Get the model which gave the highest SP index from allResults
            Establish this model as operation, to learn the following task
        end for
    end for
    Calculate the SP index average for validationResults corresponding to the TB detection task
end for
Chose the set of hyperparameters which gave the highest SP index average
```

Algorithm 2 LwF's training and evaluation using cross-validation with k-folds, k=10

```
for each itask do
    Divide dataset from task i in 10 folds
    for each ifold do
        Separate ifold as a test fold
        for each jfold different from ifold do
            Separate jfold as a validation fold
            Train ELLA or LwF model with the other 8 folds
            Apply model to all data, except for ifold, and save the results in validationResults
        end for
        Get the model which gave the highest SP index from validationResults
        Apply this model to ifold and save the results in testResults
        Apply this model to all dataset and save the results in allResults
        for each jtask < itask (evaluate performance in old tasks) do
            apply this model to all dataset from jtask and save the results in oldTasksResults
        end for
    end for
    Calculate the confidence interval at level 95% for new task with data from testResults
    Calculate the confidence interval at level 95% for model performance in old tasks with data from oldTasksResults
    Get the model which gave the highest SP index from allResults
    Establish this model as operation, to learn the following task
end for
```

We wanted to observe, with the results, 1) whether the models "forget" the previous task of pneumonia detection when it learns the new task of TB detection and 2) the difference in models performance when they learn only to detect TB and when they learn it after pneumonia detection, in the LML paradigm.

In order to analyze the first performance evaluation aspect in ELLA and LwF, the performances in pneumonia detection before and after learning to detect TB were compared. The whole dataset corresponding to the pneumonia task was used and the models trained in Algorithm 2 were applied to calculate the 95% CIs for the performance indexes. The results are displayed in Table 1.

Note that the results obtained are biased because data used as train and validation were now used in classification. This approach was chosen because only one out of the 10 models which were trained for pneumonia detection was set as the operation model, and submitted to learn TB detection. In consequence, only one out of the 10 folds could be evaluated as test, and this limitation would also introduce bias. So the obtained indexes were used to give an indication about the forgetting process of the model, but it is not possible to use them to indicate the models' performance in the task because of the bias introduced. In order to analyse the forgetting process without bias, a recommended approach would be to apply the trained models to a different pneumonia dataset.

In order to analyze the second evaluation aspect, the models' performance in TB detection when it was the only task and when it was learned after pneumonia detection were compared. When the ELLA and LwF models learn only one task, they

are equivalent to a regular logistic regression and a CNN with AlexNet design, respectively. The 95% CIs for the performance indexes were calculated. Each model was applied to the corresponding test set. The results are in Table 2.

Table 1: Results for pneumonia detection with ELLA and LwF before and after learning 2nd task, applying models to the whole dataset.

Algorithm	Index	Before learning 2nd Task	After learning 2nd Task
ELLA	Sensitivity	75.2 ± 1.2	76.0 ± 0.6
ELLA	Specificity	73.2 ± 1.3	72.5 ± 1.1
ELLA	SP Index	74.2 ± 0.5	74.2 ± 0.5
LwF	Sensitivity	87.1 ± 1.0	91.0 ± 1.3
LwF	Specificity	84.3 ± 0.8	77.7 ± 2.6
LwF	SP Index	85.7 ± 0.4	84.2 ± 0.8

Table 2: Results for TB detection with ELLA and LwF, when it is the only task learned and when it is learned after pneumonia detection, applying models to the train set.

Algorithm	Index	Learning only this task	After learning pneumonia detection
ELLA	Sensitivity	78.7 ± 12.4	77.9 ± 16.1
ELLA	Specificity	86.4 ± 9.0	86.1 ± 9.8
ELLA	SP Index	82.4 ± 6.1	81.9 ± 8.5
LwF	Sensitivity	78.3 ± 6.3	83.4 ± 4.1
LwF	Specificity	86.6 ± 7.1	86.5 ± 5.7
LwF	SP Index	82.3 ± 3.3	84.9 ± 3.1

6. Discussion

In this study, both ELLA and LwF were able to retain knowledge from the previous task after a new task was learned. For ELLA, there was no significant difference in the performance indexes obtained in pneumonia detection before and after learning TB detection. For LwF, sensitivity increased, specificity decreased and there was a small decrease in the SP index obtained in pneumonia detection after learning TB detection, at a level of confidence of 95% (Table 1).

A possible explanation for the decrease in specificity for LwF is that the model learns different abnormalities when it learns to detect TB and may transfer this knowledge to detect pneumonia, but this also may lead to more false positives. LwF presented better performance in pneumonia detection than ELLA, both before and after learning TB detection, as it was associated to higher sensitivity, specificity and SP index at a level of confidence of 95% (Table 1).

Both ELLA and LwF could learn a new task, as Table 2 showed there was no significant difference in performance at a level of confidence of 95% when they learned only TB detection and when it was learned after pneumonia detection.

The WHO established 90% of sensitivity and 70% of specificity as the performance parameters of a target product profile for TB triage [20]. In order to achieve such target operational values, corresponding detection thresholds should be determined. Our models attained specificity levels in TB detection above the target values for proper operation. This indicates that another threshold could be chosen in order to obtain a higher sensitivity and get close to the performance parameters indicated by WHO. To illustrate this, Figure 1 shows the Receiver Operating Characteristic Curve (ROC curve) [47] corresponding to the adjusted LwF models. The average curve exhibits an operation point that attains 90% of sensitivity and 70% of specificity.

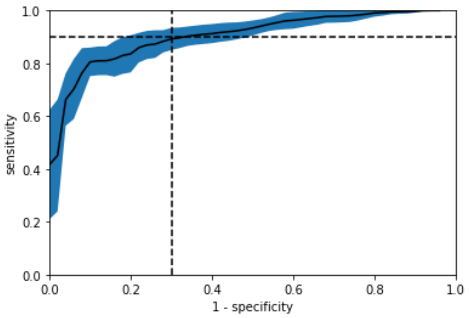

Figure 1: ROC curve for the LwF models. The black curve corresponds to the sensitivity average of the 10 adjusted models, for the specificity defined in x axis. This sensitivity was linearly interpolated for each of the models. The blue area corresponds to the error bar computed from one root mean square spread of the 10 models.

Large CIs were obtained for a level of confidence of 95% for both ELLA and LwF, but CIs from ELLA were larger (Table 2). This indicates that models adjusted in cross-validation have varying performances. This is also a consequence of having a smaller dataset for TBm comparing to the pneumonia dataset.

There was no significant difference at a level of confidence of 95% between ELLA and LwF performances for TB detection when it was the only learned task. There was also no significant difference between ELLA and LwF when TB detection was learned after pneumonia detection.

7. Conclusion

This work aimed to evaluate the potential of using LML in disease detection through CXR, verifying if two algorithms, ELLA and LwF, could retain knowledge about pneumonia detection after having learned TB detection. This ability could be verified for both algorithms.

In order to evaluate the models' knowledge retention, their performance in pneumonia detection before and after learning TB detection was compared. The results indicated that knowledge could be retained and that LwF had a better performance than ELLA.

For the evaluation about the models' ability to learn a new task, the performances of the models that learned only TB detection were compared to the models that learned it after pneumonia detection, for both ELLA and LwF. There was no significant difference in performance, which indicates that the models were able to learn the new task. Also, there was no significant difference between LwF's and ELLA's performances in detecting TB.

As future work, more diseases may be presented to the models to evaluate their knowledge retention and learning capability in a more defying and realistic scenario. Also, the models should be applied to different pneumonia and TB datasets, not involved in the training process, to better evaluate the knowledge retention.

The study had some limitations. Datasets have low quality clinical information, which hampers the validation process. Despite this limitation, we believe CADs can be trained to learn more than one task, which will be interesting for the development of more applicable tools in clinical practice.

8 Acknowledgments

This work was carried out with the support of Coordenacao de Aperfeicoamento de Pessoal de Nivel Superior Brasil (CAPES) - Financing Code 001. The authors would also like to thank FAPERJ and CNPq for supporting this work.

References

[1] World Health Organization. *The top 10 causes of death*. Number World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, Accessed on: 2022-02-27.

[2] C. W. M. Ong, G. B. Migliori, M. Raviglione, G. MacGregor-Skinner, G. Sotgiu, J.-W. Alffenaar, S. Tiberi, C. Adlhoch, T. Alonzi, S. Archuleta, S. Brusin, E. Cambau, M. R. Capobianchi, C. Castilletti, R. Centis, D. M. Cirillo, L. D'Ambrosio, G. Delogu, S. M. R. Esposito, J. Figueroa, J. S. Friedland, B. C. H. Ho, G. Ippolito, M. Jankovic, H. Y. Kim, S. R. Klintz, C. KÃ¶dmÃ¶n, E. Lalle, Y. S. Leo, C.-C. Leung, A.-G. MÃ¤rtson, M. G. Melazzini, S. N. Fard, P. Penttinen, L. Petrone, E. Petruccioli, E. Pontali, L. Saderi, M. Santin, A. Spanevello, R. v. Crevel, M. J. v. d. Werf, D. Visca, M. Viveiros, J.-P. Zellweger, A. Zumla and D. Goletti. "Epidemic and pandemic viral infections: impact on tuberculosis and the lung: A consensus by the World Association for Infectious Diseases and Immunological Disorders (WAidid), Global Tuberculosis Network (GTN), and members of the European Society of Clinical Microbiology and Infectious Diseases Study Group for Mycobacterial Infections (ESGMYC)". *European Respiratory Journal*, vol. 56, no. 4, October 2020. Publisher: European Respiratory Society Section: Task Force Report.

[3] World Health Organization. "Global tuberculosis report 2021". 2021.

[4] World Health Organization. *Chest radiography in tuberculosis detection: summary of current WHO recommendations and guidance on programmatic approaches*. World Health Organization, 2016. Number: WHO/HTM/TB/2016.20.

[5] Z. Z. Qin, T. Naheyan, M. Ruhwald, C. M. Denkinger, S. Gelaw, M. Nash, J. Creswell and S. V. Kik. "A new resource on artificial intelligence powered computer automated detection software products for tuberculosis programmes and implementers". *Tuberculosis (Edinburgh, Scotland)*, vol. 127, pp. 102049, March 2021.

[6] World Health Organization. "WHO consolidated guidelines on tuberculosis Module 2: Systematic screening for tuberculosis disease", 2021.

[7] J. O'Grady, S. V. Kik and G. Ferrara. "Extra-pulmonary tuberculosis and Xpert MTB/RIF: all about meta-analyses?" *The International Journal of Tuberculosis and Lung Disease: The Official Journal of the International Union Against Tuberculosis and Lung Disease*, vol. 19, no. 3, pp. 254, March 2015.

[8] M. Harris, A. Qi, L. Jeagal, N. Torabi, D. Menzies, A. Korobitsyn, M. Pai, R. R. Nathavitharana and F. Ahmad Khan. "A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest x-rays for pulmonary tuberculosis". *PloS One*, vol. 14, no. 9, pp. e0221339, 2019.

[9] S. Kulkarni and S. Jha. "Artificial Intelligence, Radiology, and Tuberculosis: A Review". *Academic Radiology*, vol. 27, no. 1, pp. 71–75, January 2020.

[10] E. J. Hwang, S. Park, K.-N. Jin, J. I. Kim, S. Y. Choi, J. H. Lee, J. M. Goo, J. Aum, J.-J. Yim, J. G. Cohen, G. R. Ferretti and C. M. Park. "Development and Validation of a Deep LearningâBased Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs". *JAMA Network Open*, vol. 2, no. 3, pp. e191095–e191095, March 2019. Publisher: American Medical Association.

[11] S. Sathitratanacheewin, P. Sunanta and K. Pongpirul. "Deep learning for automated classification of tuberculosis-related chest X-Ray: dataset distribution shift limits diagnostic performance generalizability". *Heliyon*, vol. 6, no. 8, pp. e04614, August 2020.

[12] World Health Organization. "WHO fact sheet on pneumonia". *https://www.who.int/news-room/fact-sheets/detail/pneumonia*.

[13] P. Ruvolo and E. Eaton. "ELLA: An Efficient Lifelong Learning Algorithm". In *International Conference on Machine Learning*, pp. 507–515. PMLR, February 2013. ISSN: 1938-7228.

[14] Z. Li and D. Hoiem. "Learning without Forgetting". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, December 2018. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[15] D. L. Silver, Q. Yang and L. Li. "Lifelong Machine Learning Systems: Beyond Learning Algorithms". In *2013 AAAI Spring Symposium Series*, March 2013.

[16] A. C. Nachiappan, K. Rahbar, X. Shi, E. S. Guy, E. J. Mortani Barbosa, G. S. Shroff, D. Ocazionez, A. E. Schlesinger, S. I. Katz and M. M. Hammer. "Pulmonary Tuberculosis: Role of Radiology in Diagnosis and Management". *RadioGraphics*, vol. 37, no. 1, pp. 52–72, January 2017. Publisher: Radiological Society of North America.

[17] A. S. Becker, C. BlÃ¼thgen, V. D. Phi van, C. Sekaggya-Wiltshire, B. Castelnuovo, A. Kambugu, J. Fehr and T. Frauenfelder. "Detection of tuberculosis patterns in digital photographs of chest X-ray images using Deep Learning: feasibility study". *The International Journal of Tuberculosis and Lung Disease: The Official Journal of the International Union Against Tuberculosis and Lung Disease*, vol. 22, no. 3, pp. 328–335, 2018.

[18] T. Naheyan. *AI Products for Tuberculosis Healthcare*. Available at: https://www.ai4hlth.org, acessed on: 2022-02-01.

[19] Z. Z. Qin, S. Ahmed, M. S. Sarker, K. Paul, A. S. S. Adel, T. Naheyan, R. Barrett, S. Banu and J. Creswell. "Can artificial intelligence (AI) be used to accurately detect tuberculosis (TB) from chest X-rays? An evaluation of five AI products for TB screening and triaging in a high TB burden setting". *arXiv:2006.05509 [cs, eess, q-bio]*, May 2021. arXiv: 2006.05509.

[20] World Health Organization. "High priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting, 28-29 April 2014, Geneva, Switzerland". Technical Report WHO/HTM/TB/2014.18, World Health Organization, 2014. number-of-pages: 96.

[21] S. V. Kik, S. M. Gelaw, M. Ruhwald, R. Song, F. A. Khan, R. v. Hest, V. Chihota, N. V. Nhung, A. Esmail, A. M. C. Garfin, G. B. Marks, O. Gorbacheva, O. W. Akkerman, K. Moropane, L. T. N. Anh, K. Dheda, G. J. Fox, N. Marano, K. LÃ¶nnroth, F. Cobelens, A. Benedetti, P. Dewan, S. Ongarello and C. M. Denkinger. "Diagnostic accuracy of chest X-ray interpretation for tuberculosis by three artificial intelligence-based software in a screening use-case: an individual patient meta-analysis of global data". Technical report, medRxiv, January 2022. Type: article.

[22] S. Meraj, R. Yaakob, A. Azman, S. N. M. Rum and A. Nazri. "Artificial Intelligence in Diagnosing Tuberculosis: A Review". 2019.

[23] A. Krizhevsky, I. Sutskever and G. E. Hinton. "ImageNet classification with deep convolutional neural networks". *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[24] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. WÃ¡ng, P.-X. Lu and G. Thoma. "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases". *Quantitative Imaging in Medicine and Surgery*, vol. 4, no. 6, pp. 475–477, December 2014.

[25] S. Hwang, H.-E. Kim, J. Jeong and H.-J. Kim. "A novel approach for tuberculosis screening based on deep convolutional neural networks". p. 97852W, March 2016.

[26] P. Lakhani and B. Sundaram. "Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks". *Radiology*, vol. 284, no. 2, pp. 574–582, April 2017. Publisher: Radiological Society of North America.

[27] R. Hadsell, D. Rao, A. A. Rusu and R. Pascanu. "Embracing Change: Continual Learning in Deep Neural Networks". *Trends in Cognitive Sciences*, vol. 24, no. 12, pp. 1028–1040, December 2020.

[28] R. Quentin, O. Awosika and L. G. Cohen. "Chapter 25 - Plasticity and recovery of function". In *Handbook of Clinical Neurology*, edited by M. D'Esposito and J. H. Grafman, volume 163 of *The Frontal Lobes*, pp. 473–483. Elsevier, January 2019.

[29] W. C. Abraham and A. Robins. "Memory retention: the synaptic stability versus plasticity dilemma". *Trends in Neurosciences*, vol. 28, no. 2, pp. 73–78, February 2005.

[30] H.-E. Kim, S. Kim and J. Lee. "Keep and learn: Continual learning by constraining the latent space for knowledge preservation in neural networks". *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11070 LNCS, pp. 520–528, 2018.

[31] A. Patra and J. A. Noble. "Incremental Learning of Fetal Heart Anatomies Using Interpretable Saliency Maps". In *Medical Image Understanding and Analysis*, edited by Y. Zheng, B. M. Williams and K. Chen, Communications in Computer and Information Science, pp. 129–141, Cham, 2020. Springer International Publishing.

[32] D. Maltoni and V. Lomonaco. "Continuous learning in single-incremental-task scenarios". *Neural Networks*, vol. 116, pp. 56–73, August 2019.

[33] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu and R. Hadsell. "Progressive Neural Networks". *arXiv:1606.04671 [cs]*, September 2016. arXiv: 1606.04671.

[34] J. Yoon, E. Yang, J. Lee and S. J. Hwang. "Lifelong Learning with Dynamically Expandable Networks". *arXiv:1708.01547 [cs]*, June 2018. arXiv: 1708.01547.

[35] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran and R. Hadsell. "Overcoming catastrophic forgetting in neural networks". *arXiv:1612.00796 [cs, stat]*, January 2017. arXiv: 1612.00796.

[36] F. Zenke, B. Poole and S. Ganguli. "Continual Learning Through Synaptic Intelligence". *arXiv:1703.04200 [cs, q-bio, stat]*, June 2017. arXiv: 1703.04200.

[37] S.-A. Rebuffi, A. Kolesnikov, G. Sperl and C. H. Lampert. "iCaRL: Incremental Classifier and Representation Learning". *arXiv:1611.07725 [cs, stat]*, April 2017. arXiv: 1611.07725.

[38] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio. "Generative Adversarial Networks". *arXiv:1406.2661 [cs, stat]*, June 2014. arXiv: 1406.2661.

[39] H. Shin, J. K. Lee, J. Kim and J. Kim. "Continual Learning with Deep Generative Replay". *arXiv:1705.08690 [cs]*, December 2017. arXiv: 1705.08690.

[40] S. Sharma, B. Maycher and G. Eschun. "Radiological imaging in pneumonia: recent innovations". *Current Opinion in Pulmonary Medicine*, vol. 13, no. 3, pp. 159–169, May 2007.

[41] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren and A. Y. Ng. "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison". In *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pp. 590–597. AAAI Press, 2019.

[42] H. Abdi and L. J. Williams. "Principal component analysis". *WIREs Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.101.

[43] A. Mallya and S. Lazebnik. "PackNet: Adding Multiple Tasks to a Single Network by Iterative Pruning". pp. 7765–7773, 2018.

[44] H. Robbins and S. Monro. "A Stochastic Approximation Method". *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, September 1951. Publisher: Institute of Mathematical Statistics.

[45] A. dos Anjos, R. C. Torres, J. M. Seixas, B. C. Ferreira and T. C. Xavier. "Neural triggering system operating on high resolution calorimetry information". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 559, no. 1, pp. 134–138, April 2006.

[46] I. Guyon, A. Saffari, G. Dror and G. Cawley. "Model Selection: Beyond the Bayesian/Frequentist Divide". *Journal of Machine Learning Research*, vol. 11, no. 3, pp. 61–87, 2010.

[47] T. D. Hoo ZH, Candlish J. "What is an ROC curve?" *Emergency Medicine Journal*, , no. 34, pp. 357–359, 2017.