Continual learning in chest radiographs of active tuberculosis and pneumonia using images generated by adversarial networks

Regina Reis da Costa Alves
Biomedical Engineering Program, COPPE.
Universidade Federal do Rio de Janeiro
Rio de Janeiro, RJ, Brazil 21941-901
Email: regina.alves@poli.ufrj.br

Frederico Caetano Jandre de Assis Tavares
Biomedical Engineering Program, COPPE.
Universidade Federal do Rio de Janeiro
Rio de Janeiro, RJ, Brazil 21941-901
Email: jandre@peb.ufrj.br

José Manoel de Seixas
Signal Processing Lab, COPPE.
Universidade Federal do Rio de Janeiro
Rio de Janeiro, RJ, Brazil 21941-901
Email: seixas@lps.ufrj.br

Otto Tavares Nascimento
Signal Processing Lab, COPPE.
Universidade Federal do Rio de Janeiro
Rio de Janeiro, RJ, Brazil 21941-901
Email: otavares93@gmail.com

João Victor da Fonseca Pinto
Signal Processing Lab, COPPE.
Universidade Federal do Rio de Janeiro
Rio de Janeiro, RJ, Brazil 21941-901
Email: joao.victor.da.fonseca.pinto@cern.ch

Anete Trajman
Faculdade de Medicina.
Universidade Federal do Rio de Janeiro
Rio de Janeiro, RJ, Brazil 21941-901
Email: atrajman@gmail.com

***Abstract*—Lower respiratory infections, including tuberculosis (TB) and pneumonia, rank among the top 10 leading causes of death worldwide. Chest radiographs (CXRs) are recommended as a screening and triage tool and computer-aided detection (CAD) softwares are an alternative to analyzing CXR. Continual learning (CL) is an option to obtain models that can identify multiple diseases by continuously learning a diverse range of radiological signs associated with each disease. In this work, we tested a CL model, Learning Without Forgetting, in learning pneumonia and TB detection using synthetic images of TB to enlarge the dataset, produced by two different Generative Adversarial Networks (GANs) and incorporated in the training process using different approaches. After learning TB detection, the model's performance in pneumonia detection has improved. Also, a potential improvement in TB detection was observed when synthetic data was used to fine tune the fully-connected layers of the model.**

I. Introduction

Lower respiratory infections rank among the top 10 leading causes of death worldwide [1]. Tuberculosis (TB) and pneumonia, including pneumonia resulting from SARS-CoV-2 infection, are listed among the most frequent lower respiratory infections [2].

TB is a contagious disease and the leading cause of death from a single infectious agent, ranking above HIV/AIDS and malaria [3].

Pneumonia is a form of acute respiratory infection that affects the lungs, specifically the alveoli, and may be caused by several infectious agents, including viruses, bacteria and fungi. Gas exchange may be impaired by fluid filling alveoli or thickening of the blood-gas barrier. Pneumonia is the single largest infectious cause of death in children worldwide [4].

Although Chest radiographs (CXRs) are a recommended tool for screening and evaluating diseases of the thorax [5], their use is limited due to a lack of radiologists in many high burden countries [6]. Costs for the health system and the patient are also barriers to the use of CXR in many of these countries. Since March 2021, the WHO included the use of computer-aided detection (CAD) software in the TB Screening Guidelines as an alternative to analyzing digital CXR for TB screening and triage in individuals aged 15 years old and above [7].

However, most available CADs' output is the probability of having TB or not, while a clinician needs to know the cause of the cough or fever (or any other symptom that could be diagnosed by the CXR) of the patient, not only if TB is probable or not. Similarities in some radiological signs among TB, pneumonia and other lung diseases highlights the importance of training a CAD model to distinguish between multiple diseases. Furthermore, considering the diverse range of radiological signs associated with each disease, it is crucial for the CAD model to continuously learn to detect them.

The paradigm of continual learning (CL) is an option to achieve these goals, since different tasks can be learned while keeping the efficiency obtained in the previous tasks. The acquired knowledge can also be used to obtain better performance in new tasks [8]. Each task can be related, for example, to a different disease detection. The learning process

can be guided in stages to obtain greater efficiency, using feedback from doctors.

A common limitation in training CAD models for healthcare applications is the availability of labeled data [9]. These models often consist of numerous parameters that need to be adjusted, requiring a large quantity of training data. To address this challenge, strategies can be employed to enlarge the dataset.

We have previously shown in [10] that two CL models, Learning without Forgetting (LwF) and Efficient Lifelong Learning Algorithm (ELLA), were able to retain knowledge about pneumonia detection and were also able to learn TB detection. In [11], synthetic images of TB were produced by two different Generative Adversarial Networks (GANs) to enlarge the dataset and were incorporated in the training process of LwF using only one approach. In this work, we use different approaches for incorporating these images to the training. The main goal was to measure whether the performance in TB detection improves as compared to using only real data, while retaining a good performance in pneumonia detection, and if a different approach can improve the previously obtained results.

The paper is structured as follows: Section II illustrates the CL paradigm by explaining diverse implementation strategies. Section III explores CL applications in the field of Medicine. Section IV provides a description of the LwF algorithm used in the experiments, along with an overview of related extended works. Section V details the conducted experiments, including pseudo-algorithms and the employed data. Section VI analyzes the achieved results from the experiments. Finally, section VII concludes the paper.

II. Continual Learning

The CL paradigm can be applied, in practice, by using different methods. For better understanding, they can be categorized into three main groups: replay, regularization-based and parameter isolation[12]. Some strategies may be a mix of different methods or may use a method that does not fit into either of the groups.

Replay methods involve either storing samples or generating synthetic samples from previous tasks. They are then presented to the model during the learning of new tasks, in order to reduce forgetting. An important method in this family is the iCarl [13], which learns new classes incrementally and stores samples from previous classes to use during training.

Regularization-based methods add a penalty term in the loss function against deviations that could lead to poor performance in previous tasks. Two important examples are the LwF algorithm [14], which will be further explained in Section IV, and Elastic Weight Consolidation [15]. The latter penalizes changes in the model's parameters which are responsible for good performance in previous tasks. One advantage of methods in this group is that they do not require storing data from previous tasks.

Parameter isolation methods are based on the addition of task-specific parameters to the network. This approach helps alleviate the stability-plasticity dilemma, since the specific parameters enable the learning of new tasks while preserving knowledge from previous tasks. However, a drawback of these methods is that the models tend to expand over time, requiring increasingly more computational resources. An important example in this group is the progressive neural network [16].

With the advance of the CL paradigm, a specific library based on pyTorch, called Avalanche, was developed for fast prototyping, training and reproducible evaluation of CL algorithms [17].

III. Continual Learning in Medicine

In recent years, there have been significant advancements in deep learning models for medical image analysis. However, there are challenges due to changes in image data distribution caused by various factors, such as different scanner manufacturers, imaging settings and local population characteristics [18]. Applying CL in this scenario helps to handle these changes, although it can also reinforce structural biases [19]. Due to this advantage, the number of works focusing on CL applications in medicine has been growing over the past few years.

Experiments involving mammography and lung CT databases demonstrate that CL can improve disease detection performance over time. This occurs because the system can be exposed to images with varying characteristics, enabling it to progressively learn and interpret a wider range of patterns[20].

CL has been applied in many image segmentation problems, such as: segmentation of hippocampus images [21]; prostate structures and brain tumor 3D images [22]; biomedical glottis images [23], wound images [24] and cardiovascular magnetic resonance [25]. It has also been applied in physiological signal processing [26].

The advances in the CL application in medicine come along with ethical challenges [27]. For example, the models may establish wrong associations, leading to conclusions that may be dangerous to the patient's health. The success of CL application also depends on clinicians' ability to understand and make good use of the outputs of the models. Another challenge is that it can be difficult to explain the decision about a diagnosis or treatment plan, once the explainability of machine learning models is limited.

IV. Learning Without Forgetting

The LwF algorithm [14] is built upon a Convolutional Neural Network (CNN) with the Alexnet architecture [28]. The LwF network has a subset of parameters θ_s, shared among all tasks, encompassing all layers except for the last one. Additionally, there are task-specific parameters θ_o, corresponding to the last layer of the network.

When a dataset associated with a new task is introduced, it is presented to the network, and the corresponding outputs y_o from the last layer of each previous task are recorded. Subsequently, parameters θ_n specific to the new task are added to the output layer, establishing connections with all neurons in the preceding layer, and initialized with random values.

During the training of the networks, a modified loss function is employed, which penalizes both errors in the classification of the new task and deviations from the recorded outputs y_o:

$$OF = \lambda_o \mathcal{L}_{old}(y_o, \hat{y}_o) + \mathcal{L}_{new}(y_n, \hat{y}_n) + \mathcal{R}(\hat{\theta}_s, \hat{\theta}_o, \hat{\theta}_n) \quad (1)$$

Where:

$$\mathcal{L}_{old}(y_o, \hat{y}_o) = -H(y'_o, \hat{y}'_o) = -\sum_{i=1}^{l} y'^{(i)}_o log \hat{y}'^{(i)}_o \quad (2)$$

$$\mathcal{L}_{new}(y_n, \hat{y}_n) = -y_n . log \hat{y}_n \quad (3)$$

$$y'^{(i)}_o = \frac{(y^{(i)}_o)^{1/T}}{\sum_{j=1}^{l} (y^{(j)}_o)^{1/T}} \quad (4)$$

$$\hat{y}'^{(i)}_o = \frac{(\hat{y}^{(i)}_o)^{1/T}}{\sum_{j=1}^{l} (\hat{y}^{(j)}_o)^{1/T}} \quad (5)$$

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;

y_n is the vector of labels for the new task;

$\hat{y}^{(i)}_o$ is the current last layer output corresponding to the previous task i;

$y^{(i)}_o$ is the recorded last layer output corresponding to the previous task i;

1 is the number of labels;

$y'^{(i)}_o$ and $\hat{y}'^{(i)}_o$ 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.

Several works were published extending the LwF algorithm. In [29] and in [30], the authors use different CNN architectures on the task of detecting different types of cancer. They also used simple data augmentation techniques, such as rotation, shifting and cropping. In [31], a task selector network is introduced, in order to decide which of the learned tasks is more suitable to classify a new element during the model operation. The authors in [32] analyze the impact of inter-task similarity in LwF's performance and show that, when tasks have low similarity, a memory budget of 1% the size of the training data can significantly help to retain knowledge from previous tasks.

V. Application of the Learning Without Forgetting algorithm to disease classification tests

The conducted experiments are an extension of the works presented in [33], [10] and [11]. They involved the learning of two different tasks: pneumonia detection, using the Stanford public dataset [34], and tuberculosis detection, using the Shenzhen public dataset [35].

The Stanford dataset was undersampled in the same way as in [33], [10] and [11], according to the following criteria:

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

Table I shows the number of images per class that were used in the experiments.

TABLE I
NUMBER OF IMAGES PER CLASS.

Class	Quantity
Non-Tuberculosis (Shenzhen)	326
Tuberculosis (Shenzhen)	336
Normal (Stanford)	1070
Pneumonia (Stanford)	1065

Given that the LwF model is built upon a CNN, it involves a significant number of parameters that need to be learned from data. Consequently, a substantial amount of training data is required to mitigate biases in the model. Therefore, two GANs were employed to produce synthetic data for training TB classification, augmenting the collected dataset. The synthetic production in this work is an extension of the work in [36].

The first GAN is called Wasserstein's GAN (WGAN) [37]. This GAN introduces a metric that controls the training, so that the synthetic data created by the generator model are consistent with the probability density function associated with real data.

The second one is called Pix2Pix, that is a conditional GAN [38]. This GAN generates synthetic data that preserves a region of interest from a real image, while generating its surroundings. In the context of this work, the region of interest refers to a lung segmentation.

In the scope of this work, only synthetic images of TB were generated. Considering that the model needs to retain knowledge of pneumonia detection after learning TB, the performance in this task was also evaluated after the incorporation of synthetic data.

Six different experiments were made, introducing in different ways during the training process the synthesized images. Also, the models were trained without synthetic images for comparison, totalizing seven experiments, which are described below.

- 1. Training only with real images (reference);
- 2. Training with real images and fine tuning the entire model with synthetic images generated by Pix2Pix (results presented in [11]);
- 3. Training with real images and fine tuning only the fully-connected layers with synthetic images generated by Pix2Pix;
- 4. Training with real images and fine tuning the entire model with synthetic images generated by WGAN (results presented in [11]);

- 5. Training with real images and fine tuning only the fully-connected layers with synthetic images generated by WGAN;
- 6. Training simultaneously with real images and synthetic images generated by Pix2Pix, which will be called an "altogether" training;
- 7. Training simultaneously with real images and synthetic images generated by WGAN.

Figure 1 illustrates the described training scheme.

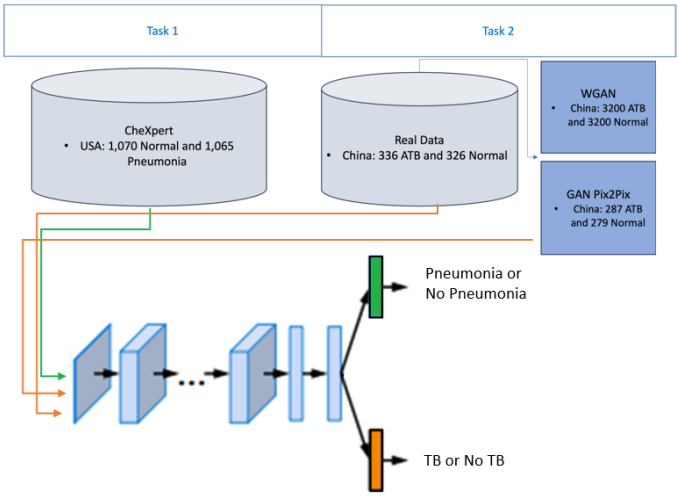

Fig. 1. Training scheme for Continual learning. Pneumonia detection is learned first and classified with the specific last layer in green. TB is learned secondly, sharing the CNN layers except for the last one, in orange, that is task-specific. The TB synthetic images are presented thirdly to finetune the model for experiments 2 to 5, and presented together with TB real data for experiments 6 and 7. Adapted from [11] and [14]

The main objective of the experiments was to analyze the impact of using synthetic data in training, and also the impact of the different configurations in which they can be introduced in the training processes.

The fine tuning stage of experiments 2 to 5 was conducted using a learning rate of 1e-4, while during the training with real images, the learning rate was of 1e-3.

The altogether training, from experiments 6 and 7, was conducted as follows: during each training epoch, batches exclusively with real images and exclusively with synthetic images were presented alternately and proportionally to the number of available images. For example, if there were 10 times more synthetic images than real images, for each 10 batches with only synthetic images, 1 batch with only real images was presented. The gradients were also weighted in order to handle the imbalance in the number of images. Using the same example, gradients from the batches with real images were weighted 10 times more than gradients from the batches with synthetic images.

For all the experiments, the cross-validation method with stratified k-folds technique [39] was used, with k=10. This method gives an estimate of the uncertainty in the results due to a limited sample. Also, for all the experiments, the operating point during TB detection training was set so that the sensitivity in the validation set was 90%, given the minimal accuracy for TB triage defined by the WHO of 90% sensitivity and 70% specificity [40]. The steps for training with only real images are described in Algorithm 1.

Algorithm 1 LwF's training and evaluation using cross-validation with k-folds, k=10 - only real images

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 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 all dataset and save the results 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

The production of synthetic data, for both Pix2Pix and WGAN, used the same 10-fold data partition employed to train the LwF model. Otherwise, synthetic images generated from real images that belong to the test fold could be inadequately presented in the train fold for LwF, biasing the training. Algorithm 2 describes the training procedure for experiments 2 to 5 and Algorithm 3 describes it for experiments 6 and 7.

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

Get the 10 trained models that resulted from the training process described in Algorithm 1
for each imodel **do**
  Get the train, validation and test folds with real data that were used when imodel was trained
  Get set with synthetic data that were generated with the same train fold
  Finetune LwF model with the synthetic data. During training, use the validation fold with real data for validation.
end for

Table II and Table III display the sensitivity, specificity and SP index concerning the pneumonia detection task and the TB detection task, respectively. The results refer to the application of the models on the test folds. The SP index [41] is a function of the two former indexes, as shown in Equation 6.

Algorithm 3 LwF's training and evaluation using cross-validation with k-folds, k=10 - altogether training

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
      Separate the other 8 folds for training
      Get set with synthetic data that were generated with the same train fold
      Train LwF model with alterning batches of real and synthetic data
      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 all dataset and save the results 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

$$SP = \sqrt{\sqrt{sens * spec} * \frac{sens + spec}{2}} \tag{6}$$

Table II additionally presents, in its first row, the results when the model has learned only pneumonia detection. Other rows present the results when the model has already learned TB detection, for experiments 1 to 7.

Figures 2 and 3 present the same results as Table I and II, but only for the SP Index, in a graphical manner.

In order to evaluate the statistical significance of the obtained results, p-values were calculated using two Hypothesis Tests. The first one evaluates whether pneumonia detection changed after learning TB detection, for the 7 experiments:

H0: The SP Index mean for pneumonia detection after learning TB is not different from the SP Index mean before learning TB.

H1: The SP Index mean for pneumonia detection after learning TB is different from the SP Index mean before learning TB.

P-Value was 0.0002 for experiment 1, 0.0003 for experiment 1, 0.0005 for experiment 3 and 0.0001 for experiments 4 to 7.

The second one evaluates whether TB detection changed with the synthetic data incorporation, for the 6 experiments involving synthetic data:

H0: The SP Index mean for TB detection after incorporating synthetic data is not different from SP Index mean only with real data (experiment 1 - reference).

H1: The SP Index mean for TB detection after incorporating synthetic data is different from SP Index mean only with real data.

P-Value was 0.9 for experiment 2, 0.65 for experiment 3, 0.97 for experiment 4, 0.0008 for experiment 5, 0.078 for experiment 6 and 0.0018 for experiment 7.

TABLE II
RESULTS FOR PNEUMONIA DETECTION WITH LWF BEFORE LEARNING AND AFTER LEARNING TB DETECTION, FOR THE 7 DESCRIBED EXPERIMENTS - APPLICATION OF THE MODELS ON THE TEST FOLDS.

Case	Sensitivity	Specificity	SP Index
Before learning TB	79.4 +- 4.1	75.0 +- 5.5	77.1 +- 2.2
Training only with TB real images (reference)	85.0 +- 6.9	84.5 +- 5.2	84.7 +- 4.8
fine tuning with pix2pix (all layers)	84.1 +- 7.0	85.0 +- 5.1	84.5 +- 4.7
fine tuning with pix2pix (only fully connected layers)	84.4 +- 7.2	84.5 +- 5.2	84.4 +- 5.0
fine tuning with WGAN (all layers)	83.4 +- 5.3	84.9 +- 5.6	84.1 +- 3.7
fine tuning with WGAN (only fully connected layers)	89.1 +- 5.7	77.2 +- 4.3	83.0 +- 2.8
Training altogether, real images + pix2pix	85.7 +- 6.9	84.4 +- 3.6	85.0 +- 4.6
Training altogether, real images + WGAN	88.4 +- 6.5	81.8 +- 4.6	85.0 +- 4.5

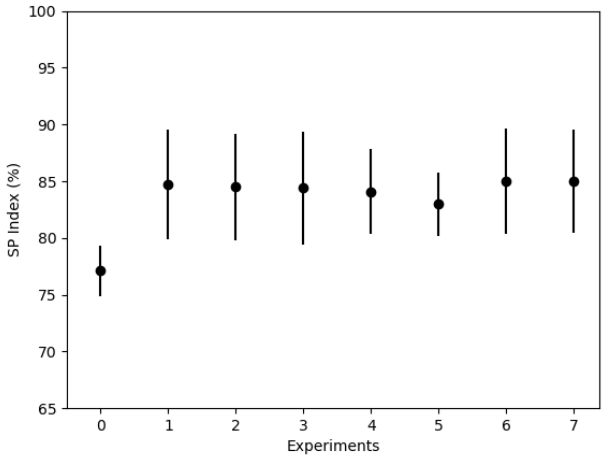

Fig. 2. SP Indexes for pneumonia detection with LwF, before and after learning TB detection, across the seven described experiments. Dots indicate mean SP Index; lines show ±1 standard deviation.

VI. DISCUSSION

The results in Table II support the interpretation that knowledge for pneumonia detection was preserved in all experiments. Additionally, the performance in pneumonia detection was enhanced after training for TB detection in all experiments. This can be more clearly observed in Figure 2, that shows that SP Index for experiment 0 (before learning TB) is below the SP Indexes for experiments 1-7. The first Hypothesis

TABLE III
RESULTS FOR TB DETECTION WITH LWF FOR THE 7 DESCRIBED EXPERIMENTS - APPLICATION OF THE MODELS ON THE TEST FOLDS.

Experiment	Sensitivity	Specificity	SP Index
Training only with TB real images (reference)	89.3 +- 4.8	71.8 +- 9.5	80.2 +- 4.9
fine tuning with pix2pix (all layers)	88.4 +- 5.0	72.2 +- 11.9	79.9 +- 5.6
fine tuning with pix2pix (only fully connected layers)	89.6 +- 4.6	69.7 +- 12.5	79.1 +- 5.8
fine tuning with WGAN (all layers)	90.8 +- 6.2	70.6 +- 14.1	80.1 +- 6.7
fine tuning with WGAN (only fully connected layers)	89.0 +- 5.5	85.5 +- 7.1	87.1 +- 2.3
Training altogether, real images + pix2pix	91.4 +- 6.2	59.8 +- 16.8	74.2 +- 8.9
Training altogether, real images + WGAN	87.5 +- 7.3	57.9 +- 15.0	71.4 +- 5.8

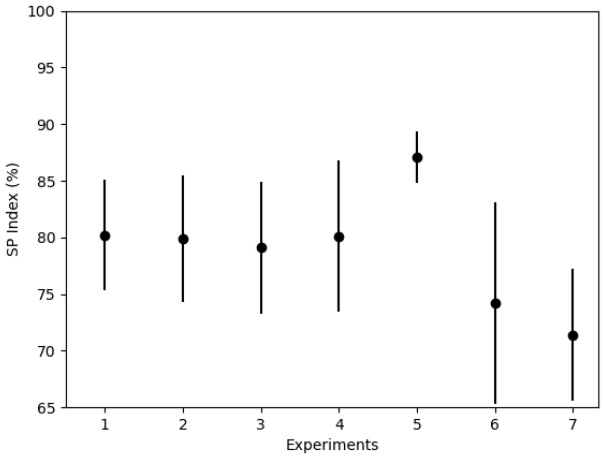

Fig. 3. SP Indexes for TB detection with LwF across the seven described experiments. Dots indicate mean SP Index; lines show ±1 standard deviation.

Test suggests very strong evidence of this enhancement, as p-values are lower than 0.001 [42].

These findings indicate that the knowledge acquired for TB detection positively influenced pneumonia detection. One potential explanation, to be probed further, is that there are similarities between those tasks that enable improvements in one by training for the other.

Considering the results displayed in Table III, the mean SP Index for TB detection is higher than the reference case only for experiment 5. However, the second Hypothesis Test suggests little or no evidence that experiments 2 to 4 result in poorer outcomes in TB detection, given that the p-value is above 0.1. The p-values associated with the experiments with the altogether training (6 and 7) indicate weak evidence and strong evidence, respectively, for inferior outcomes in TB detection.

On the other hand, fine tuning with synthetic images generated by WGAN, when the convolutional layers were freezed during fine tuning (experiment 5), resulted in a SP index mean 7% higher than the reference. Figure 3 highlights the superior results achieved in Experiment 5, with a higher mean and lower deviation. Additionally, the p-value lower than 0.001 indicates very strong evidence for an improvement in TB detection, which indicates a promising technique to be incorporated on the training.

VII. CONCLUSION

This work presented an application of the CL paradigm to pneumonia and TB detection through CXR images, using the LwF algorithm. GANs were employed to produce synthetic images for TB task. These synthetic images were incorporated to the training by using different approaches, in six experiments. Additionally, a reference experiment was conducted using only real images, resulting in a total of seven experiments.

After learning TB detection, the model's performance in pneumonia detection improved, which hints at the algorithm's capability of sharing knowledge between tasks, as previously shown in [10] and [11].

The incorporation of synthetic data showed a potential to improve the model's performance, but not with all approaches. A potential improvement was observed when synthetic data was used to finetune the fully-connected layers of the model. However, poor results were obtained when they were incorporated in training together with the real data ("altogether training").

This is a promising result for improving the clinical utility of CADs. Proposals for future work include hyperparameter optimization in the presented approaches, aiming at improving results obtained especially in the "altogether" training. Also, synthetic data may be produced and utilized also in the pneumonia detection task, which is expected to also enhance performance of the model, with special interest in the assessment of potential mutual benefits for both detection tasks.

VIII. ACKNOWLEDGMENTS

The authors thank FAPERJ and CNPq (project 440129/2020-6) for partial support to this work. This work was carried out with the support of Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Brasil (CAPES) - Código de Financiamento 001.

REFERENCES

[1] D. R. Silva, F. C. d. Q. Mello, L. D'Ambrosio, R. Centis, M. P. Dalcolmo and Migliori. "Tuberculose e COVID-19, o novo dueto maldito: quais as diferenças entre Brasil e Europa?" vol. 47, no. 2, pp. e20210044–e20210044, 2021.

[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. Ködmön, E. Lalle, Y. S. Leo, C.-C. Leung, A.-G. Märtson, 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 2022: World Health Organization*. 2022.

[4] World Health Organization. "Fact sheets: Pneumonia in Children". Technical report, World Health Organization, 2022.

[5] G. Frija, I. Blažić, D. P. Frush, M. Hierath, M. Kawooya, L. Donoso-Bach and B. Brkljačić. "How to improve access to medical imaging in low- and middle-income countries ?" *EClinicalMedicine*, vol. 38, pp. 101034, August 2021.

[6] 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.

[7] World Health Organization. "WHO consolidated guidelines on tuberculosis Module 2: Screening – Systematic screening for tuberculosis disease". *Geneva, Switzerland*, 2021.

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

[9] Z. Eaton-Rosen, F. Bragman, S. Ourselin and M. J. Cardoso. "Improving Data Augmentation for Medical Image Segmentation". June 2018.

[10] R. Alves, F. Tavares, J. Seixas and A. Trajman. "Applying the Lifelong Machine Learning Paradigm in Tuberculosis Triage". *Learning and NonLinear Models*, vol. 20, no. 2, pp. 63–73, 2022.

[11] O. Tavares, R. Alves, J. Seixas, F. Jandre and A. Trajman. "An Innovative Artificial Intelligence Approach For Fighting Against Tuberculosis". In *The 2nd BRICS Postgraduate Forum*, 2023.

[12] M. Delange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh and T. Tuytelaars. "A Continual Learning Survey: Defying Forgetting in Classification Tasks". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, pp. 1–1, February 2021.

[13] 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.

[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] 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.

[16] 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.

[17] V. Lomonaco, L. Pellegrini, A. Cossu, A. Carta, G. Graffieti, T. L. Hayes, M. D. Lange, M. Masana, J. Pomponi, G. van de Ven, M. Mundt, Q. She, K. Cooper, J. Forest, E. Belouadah, S. Calderara, G. I. Parisi, F. Cuzzolin, A. Tolias, S. Scardapane, L. Antiga, S. Amhad, A. Popescu, C. Kanan, J. van de Weijer, T. Tuytelaars, D. Bacciu and D. Maltoni. "Avalanche: an End-to-End Library for Continual Learning". In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2nd Continual Learning in Computer Vision Workshop, 2021.

[18] K. Shu, H. Li, J. Cheng, Q. Guo, L. Leng, J. Liao, Y. Hu and J. Liu. "Replay-Oriented Gradient Projection Memory for Continual Learning in Medical Scenarios". In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1724–1729, December 2022.

[19] J. E. Carolan, J. McGonigle, A. Dennis, P. Lorgelly and A. Banerjee. "Technology-Enabled, Evidence-Driven, and Patient-Centered: The Way Forward for Regulating Software as a Medical Device". *JMIR medical informatics*, vol. 10, no. 1, pp. e34038, January 2022.

[20] C. Muramamatsu, M. Nishio, M. Oiwa, M. Yakami, T. Kubo and H. Fujita. "Investigation on continual training of computer-aided diagnosis systems by semi-supervised learning". In *2022 4th International Conference on Intelligent Medicine and Image Processing*, IMIP 2022, pp. 58–62, New York, NY, USA, April 2022. Association for Computing Machinery.

[21] A. Ranem, C. González and A. Mukhopadhyay. "Continual Hippocampus Segmentation with Transformers". pp. 3710–3719. IEEE Computer Society, June 2022.

[22] M. Tian, Q. Yang and Y. Gao. "Multi-scale Multi-task Distillation for Incremental 3D Medical Image Segmentation". In *Computer Vision – ECCV 2022 Workshops*, edited by L. Karlinsky, T. Michaeli and K. Nishino, Lecture Notes in Computer Science, pp. 369–384, Cham, 2023. Springer Nature Switzerland.

[23] R. Groh, S. Dürr, A. Schützenberger, M. Semmler and A. M. Kist. "Long-term performance assessment of fully automatic biomedical glottis segmentation at the point of care". *PLOS ONE*, vol. 17, no. 9, pp. e0266989, September 2022. Publisher: Public Library of Science.

[24] N. Curti, Y. Merli, C. Zengarini, E. Giampieri, A. Merlotti, D. Dall'Olio, E. Marcelli, T. Bianchi and G. Castellani. "Effectiveness of Semi-Supervised Active Learning in Automated Wound Image Segmentation". *International Journal of Molecular Sciences*, vol. 24, no. 1, pp. 706, January 2023. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[25] S. Chen, D. An, C. Feng, Z. Bian and L.-M. Wu. "Segmentation of Pericardial Adipose Tissue in CMR Images: a Benchmark Dataset MRPEAT and a Triple-Stage Network 3SUnet". *IEEE transactions on medical imaging*, vol. PP, March 2023.

[26] L. Sun, J. Wu, Y. Xu and Y. Zhang. "A federated learning and blockchain framework for physiological signal classification based on continual learning". *Information Sciences*, vol. 630, pp. 586–598, June 2023.

[27] J. Hatherley and R. Sparrow. "Diachronic and synchronic variation in the performance of adaptive machine learning systems: the ethical challenges". *Journal of the American Medical Informatics Association*, vol. 30, no. 2, pp. 361–366, February 2023.

[28] 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.

[29] M. Subramanian, J. Cho, V. E. Sathishkumar and O. S. Naren. "Multiple Types of Cancer Classification Using CT/MRI Images Based on Learning Without Forgetting Powered Deep Learning Models". *IEEE Access*, vol. 11, pp. 10336–10354, 2023. Conference Name: IEEE Access.

[30] G. Oren and L. Wolf. "In Defense of the Learning Without Forgetting for Task Incremental Learning". In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2209–2218, Montreal, BC, Canada, October 2021. IEEE.

[31] N. Ammour, H. Alhichri, Y. Bazi and N. Alajlan. "LwF-ECG: Learning-without-forgetting approach for electrocardiogram heartbeat classification based on memory with task selector". *Computers in Biology and Medicine*, vol. 137, pp. 104807, October 2021.

[32] A. El Khatib, M. Nasr and F. Karray. "Accounting for the Effect of Inter-Task Similarity in Continual Learning Models". In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1241–1247, October 2021. ISSN: 2577-1655.

[33] R. Alves, F. Tavares, J. Seixas and A. Trajman. "Introducing lifelong machine learning in the active tuberculosis classification through chest radiographs". In *SBIC*, 2021.

[34] 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.

[35] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, 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.

[36] Otto Tavares, José Seixas and Anete Trajman. "Data-augmentation de dados de radiografia de torax no contexto de aprendizagem profunda". In *SBIC*, 2021.

[37] M. Arjovsky, S. Chintala and L. Bottou. "Wasserstein Generative Adversarial Networks". In *Proceedings of the 34th International Conference on Machine Learning*, pp. 214–223. PMLR, July 2017. ISSN: 2640-3498.

[38] M. Mirza and S. Osindero. "Conditional Generative Adversarial Nets", November 2014. arXiv:1411.1784 [cs, stat].

[39] 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.

[40] 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.

[41] 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.

[42] L. Held and M. Ott. "On p-Values and Bayes Factors". *Annual Review of Statistics and Its Application*, vol. 5, no. 1, pp. 393–419, 2018.