A Robust Pneumonia Classification Approach based on Self-Paced Learning

Sarpong Kwadwo Asare[1], Fei You[2], Obed Tettey Nartey[3]
School of Electronic Science and Engineering
University of Electronic Science and Technology of China
Chengdu, 611731, P.R. China

***Abstract*—This study proposes a self-paced learning scheme that integrates self-training and deep learning to select and learn labeled and unlabeled data samples for classifying anterior-posterior chest images as either being pneumonia-infected or normal. With this new approach, a model is first trained with labeled data. The model is evaluated on unlabeled data to generate pseudo labels for the unlabeled data. Using a novel selection scheme, the pseudo-labeled samples are then selected to update the model in next training iteration of the semi-supervised training process. The selected pseudo-labeled images to be added to the next training iteration are images with the most confident probabilities from every unlabeled class. Such a selection scheme prevents mistake reinforcement, which is a prevalent occurrence in self-training. With deep models having the tendency to latch onto well-represented class samples while ignoring less transferable and represented classes, especially in the case of unbalanced data, the proposed method utilizes a novel algorithm for the generation and selection of reliable top-K pseudo-labeled samples to be used in updating the model during the next training phase. Such an approach does not only force the model to learn the hard samples in the training data, it also helps enlarge the training set by generating enough samples that satisfy the hunger of deep models. Extensive experimental evaluation of the proposed method yields higher accuracy results compared to methods mentioned in the literature on the same dataset, an indication of the effectiveness of the proposed method.**

Keywords*—*Anterior-posterior chest images; self-paced learning; self-training; pneumonia classification

I. Introduction

The increasing levels of pollution in many developing countries put millions of people in such countries at the risk of contracting lung related infections. Statistics from the World Health Organization (WHO) estimates that more than four million premature deaths occur every year as a result of diseases related to pollution, which includes pneumonia [1]. A report by [2] shows that on a yearly basis, the number of people infected with pneumonia is over 150 million, with majority of these numbers coming from children below five years old. This worrying trend necessitates the automatic and accurate computer-aided detection of pneumonia in its early stages, that ultimately leads to accurate and effective diagnosis and treatment.

In recent years, machine learning and deep learning approaches are spearheading the surge in the computer-aided systems for diagnosis in the medical domain. Deep learning methods have successfully implemented in medical imaging tasks, including but not limited to classification [3] [4], detection [5] [6] and segmentation [7] [8]. Convolutional neural networks (CNN), a deep learning method, has been obtaining impressive performances in a wide range of tasks. Contributing greatly to the successes of CNN models is their inherent nature. The hierarchical layout of CNN models enable different layers to learn different features or patterns from data related to a specific task. However, an underlying attribute of CNN models is that, they require huge amounts of well-labeled data during training in order to arrive at satisfactory outcomes. The absence of such kind of data leaves the models prone to overfitting, which degrades their performances due to poor generalization.

A significant challenge with medical imaging tasks is obtaining ample labels for data samples. Moreover, for deep model to generalize well on data, a significant amount of images samples required during training. Such huge amounts of image samples are virtually non-existent in the medical domain. Compounding this problem is the process of data labeling (in the case where sufficient amounts of data exists). The data labeling process is a laborious and time-consuming one, which require expertise knowledge. To efficiently harness and maximize available data, existing methods mentioned in the literature resort to training CNN models from scratch and adopting data augmentation schemes in a bid to augment and enlarge the training set [9] [10]. The methods adopted in these works are supervised learning approaches, which typically use only labeled data. Nonetheless, an effective approach to reducing the cost of data labeling yet generating more data sample is to incorporate both labeled and unlabeled data in the training process via semi-supervised learning. Unlabeled data is rather inexpensive and abundant compared to the process of obtaining well-labeled data. This idea of using both labeled and unlabeled data in classification tasks has been less exploited in chest x-ray and pneumonia classification. The principal idea of semi-supervised learning is to utilize both labeled and unlabeled image samples in building efficient learners, instead of only using labeled image samples.

This work proposes a novel semi-supervised learning approach that utilizes self-training to classify chest x-ray images as either normal or pneumonia-infected. To this end, the proposed approach adopts self-paced learning [11], a learning paradigm inspired by the way humans learn, where a learner first learns easy samples and followed by the gradual addition of more complex samples in a meaningful way, resulting the in the learner becoming more matured and robust. To incorporate "easy-to-hard" samples into the training data, the proposed approach utilizes both labeled and unlabeled data. Pseudo-labels are assigned to the unlabeled data and target specific model is trained with pseudo-labels via self-training, as though the pseudo-labels were true labels of the unlabeled data samples. The principal idea behind self-training is generating a set of pseudo-labels which correspond to a high confidence probability score. With self-training, a model is trained with a training set that comprises the generated pseudo-labels based on the assumption that, only target samples with the highest prediction probability are selected to update the training set in the next iteration.

In the case where this assumption isn't met, a model may reinforce incorrectly labeled data into the next training iteration, and a situation known as mistake reinforcement occurs. Mistake reinforcement ultimately degrades the performance of a model. In order to prevent such a scenario from occurring, the proposed approach utilizes a novel pseudo-label selection algorithm to generate and select the top-K pseudo-labeled samples, to be used in augmenting the training set during the next training iteration. The proposed scheme forces the base learner to learn hard samples, in that, samples from both well represented and less represented classes are added to the training set. Using a simple CNN model trained from scratch as the base learner, the proposed approach yields a significantly higher accuracy compared to supervised methods mentioned in the literature.

The contributions of this work are as follows;

- A novel CNN-based self-training framework is proposed to classify anterior-posterior chest x-ray images as normal or pneumonia-infected by utilizing both labeled and unlabeled data. In this way, the machine learning technique of self-paced learning and CNN are integrated to classify chest x-ray images.
- A new heuristic pseudo-label generation and selection algorithm is proposed to generate and select the top-k most reliable pseudo-labels and their corresponding pseudo-labeled samples in updating the model, alleviating the issue of reinforcing incorrectly labeled samples in updating the CNN model, a drawback which characterizes conventional self-training.
- The proposed heuristic algorithm is capable of making the self-paced learning method jointly learn a good classifier and optimize the pseudo-labels. This is to ensure that a chunk of the pseudo-labeled samples are not ignored in the selection process at the same time solve the challenge of amassing enough reliable data for deep CNN based models.
- Finally, the problem is formulated as a loss minimization scheme that is solved by utilizing an end-to-end approach to learn a good learner and also learn the domain invariant features in chest x-ray images to distinguish pneumonia and normal tissue images.

The rest of this work is organized as follows; Section II surveys some works mentioned in the literature relating pneumonia classification, the self-paced learning scheme for pneumonia classification is introduced in Section III, with materials used and corresponding experiments described in Section IV. Section V details results and discussions and this work is concluded in Section VI.

II. Related Work

Recent advancements in deep learning methods have led to successes in many computer-aided diagnosis and medical imaging tasks including classification, segmentation and detection. Commendable results have been reported in the literature, that show exciting prospects in applying deep learning models in medical related tasks. Over the years, trend is evident in the development of several deep learning algorithms that seek to improve accuracies and minimize loss. These models have achieved excellent accuracy performances in classification tasks on natural images datasets such as the CIFAR, MNIST and ImageNet. For the particular case of examining chest x-ray images, task ranging from detecting abnormalities to classifying such abnormalities have been reported by some works [12], [13], [14], [15]. Authors in [10] classified chest x-ray images as pneumonia-infected or otherwise by using a CNN model. The authors adopted data augmentation techniques for training a CNN model, and obtained a classification of 93.73%. Authors in [16] developed CheXNet, a 121-layer CNN model that was trained on the ChestX-ray14. The authors compared the performance of CheXNet with that of radiologists, and obtained performance that exceeded that of pathologists. Performance was extended to cover all 14 diseases in the ChestX-ray14 dataset. In [17], authors used the pre-trained VGG16 model to detect and pneumonia and discriminate bacterial and viral pneumonia. Their approach focused on localizing the affected regions in an image, and reported accuracy performances of 96.2% and 93.6%. Authors in [18] perform binary classification using a CNN model on 5863 chest x-ray images to discriminate pneumonia and normal images. They reported an accuracy of 95.30% The work in [19] presented an 18-layer CNN architecture trained on 5863 chest x-ray images to perform normal versus pneumonia classification and reported an accuracy of 94.39%. Again, a deep learning method was adopted in classifying images as either normal or pneumonia in [20] with the authors reporting accuracies between 96-97%.

The methods adopted in the above-mentioned approaches rely on supervised learning, where only labeled data is used in the training process. The proposed approach adopts a semi-supervised learning approach to make use of both labeled and unlabeled data in classifying chest x-ray images as either pneumonia or normal.

III. Self-Paced Learning for Pneumonia Classification

In the medical imaging domain, the ratio of unlabeled data to labeled data presents a significant challenge in successfully accomplishing tasks. The task of obtaining well-labeled data is time-consuming, and also requires guidance from experts. These factors render such a process expensive and laborious. As such, a technique that can exploit both unlabeled and labeled data in training a CNN learner presents significant and exciting prospects in this domain. Semi-supervised learning incorporates both labeled and unlabeled data in building better learners. Semi-supervised learning algorithms have been adopted in some works mentioned in the literature for some classification tasks [21][22][23]. The core idea behind semi-supervised learning involves training a learner on labeled data and using the base learner to predict labels for unlabeled data.

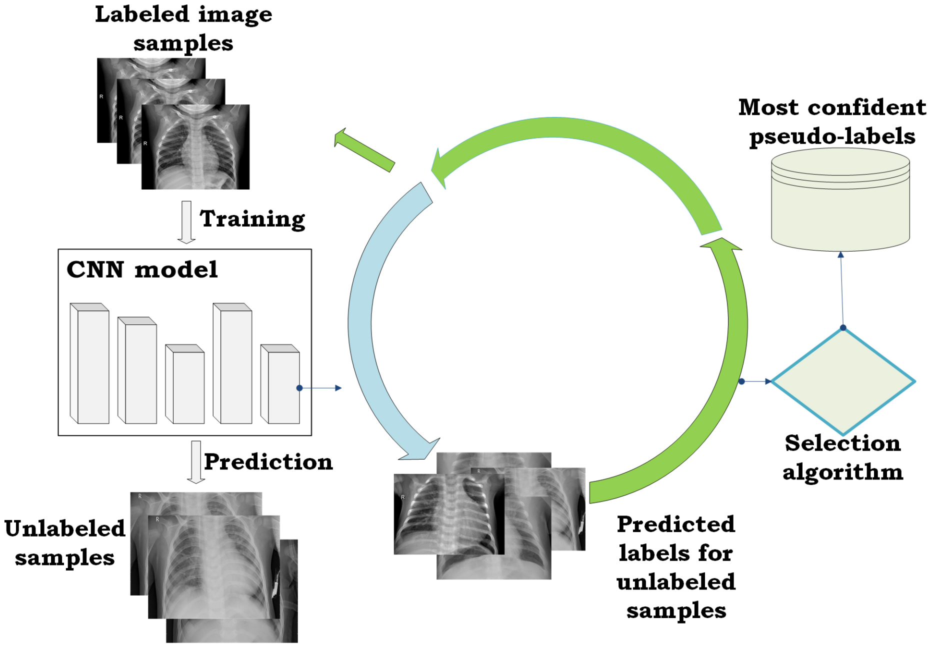

Fig. 1. Algorithm workflow of the proposed approach. A deep CNN model is first trained from scratch with labeled data samples. Pseudo-labeled samples are generated from unlabeled data samples and the most confident pseudo-labeled samples via the selection algorithm. The selected confident labels together with the labeled image samples are used as training data for the next training iteration

Nonetheless, a prevalent occurrence in many datasets is that, some classes tend to be better represented compared to others and the samples in the various classes do not always present an accurate representation of the characteristic differences between the classes themselves. When such a situation occurs, a learner tends to easily latch on to features from classes with a higher representation other than considering samples from all classes, irrespective of their representation. Ultimately, a learner literally abandons robust and versatile features relevant to its learning process, and this subsequently impacts the ability of the learner to generalize well on data.

To curb such an occurrence, a model can be gradually introduced to training or data samples in an easy-to-hard manner, utilizing a class-wise confidence probability in selecting pseudo-labels with higher confidence scores for updating the learner in the next training iteration. This is the core idea behind self-paced learning (SPL) and it has been adopted in some works [24][25][26]. It is a learning scheme that mimics the learning process of humans and animals by adding easy-to-hard samples in gradual manner. SPL has been demonstrated to be helpful in preventing bad local minima and achieving a better generalization outcome [11]. The classifier or learner determines the sequence of gradually training samples, and this is where SPL introduces a regularization term into the learning objective, enabling a learner to jointly learn a curriculum that consists of easy-to-hard or complex samples.

A review of the SPL paradigm is first introduced before introducing the proposed approach. Considering a given training data, $D = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i \in \mathbb{R}^m$ represents the ith observed sample, with y_i being its corresponding label. The loss function, which estimates the cost between the ground truth label y_i and the estimated label $f(\mathbf{x}_i, \mathbf{w})$, is denoted as $L(y_i, f(\mathbf{x}_i, \mathbf{w}))$. $\mathbf{w}$ represents model parameters inside the decision function f. SPL aims at jointly learning the model parameter $\mathbf{w}$ and the weight variable $\mathbf{v} = [v_1, ..., v_n]$ by minimizing

$$min_{\mathbf{w},\mathbf{v}}\mathbb{E}(\mathbf{w}, \mathbf{v}; \lambda) = \sum_{i=1}^{n} v_i L(y_i, f(\mathbf{x}_i, \mathbf{w})) - \lambda \sum_{i=1}^{n} v_i, \mathbf{s.t.v} \in [0,1]^n \quad (1)$$

where λ is the parameter that controls the rate at which the model learns new samples which has a direct correspondence with the "age" of the model. When the value of λ is set to very small, the model only considers "easy" samples with small losses. With an increase in the growth of λ, more samples with larger losses are gradually added, making the model mature.

With reference to minimizing the loss function, the proposed method adopts a semi-supervised model with softmax output that is solved using an end-to-end approach to learn a good classifier. This work proposes to formulate the loss function as;

$$minL_{st}\left(\mathbf{W}\right)_W = -\sum_{s=1}^{S}\sum_{n=1}^{N}\mathcal{Y}_{s,n}^{T}\log\left(P_n(W, I_s)\right) - \sum_{s=1}^{T}\sum_{n=1}^{N}\mathcal{Y}_{t,n}^{T}\log\left(P_n(W, I_t)\right). \quad (2)$$

where $\boldsymbol{I_s}$ represents the image in the source domain indexed by s $= 1, 2, 3, ..., \mathbf{S}$. $\mathcal{Y}_{s,n}$ denotes the true labels for the $\boldsymbol{nth}$ image (n = 1,2,...,N) for $\boldsymbol{I_s}$ and $\boldsymbol{W}$ represents the network weights. The softmax output containing the class probabilities is denoted as $\boldsymbol{P_n(w, I_s)}$. The definitions for

$I_t, \mathcal{Y}_{t,n}$ and $p_n(w, I_t)$ at the time of evaluation are similar. In the likelihood that some target labels are unavailable, the model presumes that these labels are hidden and learns from approximate target labels $\hat{\mathcal{Y}}$ for $\hat{\mathcal{C}}$, which indicates the number of samples.The term $\hat{\mathcal{Y}}$ (indicated in Equation 3) is referred to as the pseudo-labels to be used in the self-training scheme.

$$\begin{aligned} minL_{st}\left(W,\hat{\mathcal{Y}}\right)_{W,\hat{\mathcal{Y}}} = & -\sum_{s=1}^{S}\sum_{n=1}^{N}\mathcal{Y}_{s,n}^{T}\log\left(P_n\left(W,I_s\right)\right) \\ & -\sum_{s=1}^{T}\sum_{n=1}^{N}\hat{\mathcal{Y}}_{t,n}^{T}\log\left(P_n(W,I_t)\right). \end{aligned} \tag{3}$$

A. Self-Training with Self-Paced Learning

Conventional self-training is based on the assumption that, the high confidence predictions of a leaner are correct. Assuming an input instance x with label y, and given a learner $f : \chi \mapsto \mathcal{Y}$, labeled data $(X_l, Y_l) = \{x_{1:l}, y_{1:l}\}$, unlabeled data $X_u = \{X_{l+1:n}\}$, in self-training, the learner f is first trained from (X_l, Y_l) via supervised learning. Then the learner f is used to predict the labels for the unlabeled data X_u. A subset S, which typically comprises the few unlabeled instances X_u with the most confident predictions, is selected together with predicted labels to be added to the labeled data (X_l, Y_l). The learner is re-trained on the labeled data (which is much larger now) and the procedure is repeated. Typical of conventional self-training, an early mistake by the learner can reinforce wrong predictions into the training set for the next training iteration.

Algorithm 1 details the procedure of the self-training scheme. It starts by training a classifier with labeled samples, subsequently using the learned classifier to predict labels for non-annotated samples I_t. The predictions are known as generated pseudo-labels and with the novel selection scheme, the top-K pseudo-labeled samples are selected and added to annotated labeled set for the next model training. This process is executed iteratively until a stopping criterion is met. The fundamental idea behind the notion of an "easy-to-hard" approach is the generation of pseudo-labels from the most confident and correct predictions, updating the model with the augmented samples, and then exploring the remaining less-confident pseudo-labels. With this approach, Equation 3 is modified into;

$$\begin{aligned} minL_{st}\left(W,\hat{\mathcal{Y}}\right)_{W,\hat{\mathcal{Y}}} = & -\sum_{s=1}^{S}\sum_{n=1}^{N}\mathcal{Y}_{s,n}^{T}\log\left(P_n\left(W,I_s\right)\right) \\ & -\sum_{s=1}^{T}\sum_{n=1}^{N}\left[\int_{1}^{2}\hat{\mathcal{Y}}_{t,n}^{T}\log\left(P_n\left(W,I_t\right)\right)\right. \\ & \left. +k\left|\hat{\mathcal{Y}}_{t,n}^{T}\right|_{1}\right]. \end{aligned} \tag{4}$$

In the scenario where the pseudo-label $\hat{\mathcal{Y}}$ (in Equation 4) is ignored, the value of $\mathcal{Y}$ is assigned to zero. Again, to avoid the case of ignoring a substantial amount of pseudo-labels, L_1 regularizer is added to the loss function in Equation 4. $k > 0$ ensures the selection of more pseudo-labels during training.

In order to minimize the loss in Equation 4, 1), W is first initialize and the loss is minimized *w.r.t* $\hat{\mathcal{Y}}_{t,n}$ and then 2) $\hat{\mathcal{Y}}_{t,n}$ and the objective function is optimized *w.r.t* W. Executing step 1 and step 2 is considered to be a single iteration. In 1), optimizing discrete variables requires a non-linear function. Given that $k > 0$, the entire process in 1) can be re-expressed as;

Algorithm 1: Self-paced learning algorithm

```
input  : Deep Learning Network $P(w)$, unlabeled
         Images $I_t$, amount $K$
output: $Classifier(C)$
Train a network $P(w)$ from scratch with labeled
  samples $I_s$
for $k \leftarrow 1$ to $N$ do
    • Test and predict on unlabeled samples $I_t$;
    • Generate pseudo-labels for $I_t$ using predictions;
    • Select $K$-pseudo-labeled samples.;
    • Append $K$-pseudo-labeled samples to labeled set
      $(I_s + K(I_t))$
    • Re-train $P(w)$ on both $I_s$ and $K$-pseudo-labeled
      samples $(I_s + K(I_t))$
end
$C = updated(P(w))$;
Return $C$
```

$$\begin{aligned} & min_{\hat{\mathcal{Y}}} -\sum_{t=1}^{T}\sum_{n=1}^{N}\left[\sum\hat{\mathcal{Y}}_{t,y}^{(c)}\log\left(p_n(c|w,I_t)\right)+k|\hat{\mathcal{Y}}_{t,n}|_1\right]. \\ & s.t. \quad k > 0 \end{aligned} \tag{5}$$

The pseudo-labels ought to satisfy one of the following conditions; 1) either it is a discrete one-hot vector or 2) a vector with a null magnitude. As such, the pseudo-label framework is optimized via;

$$\hat{\mathcal{Y}}_{t,y}^{(c*)} = \begin{cases} 1, & \text{if } c = \arg\max p_n\left(c|w,I_t\right), \\ & p_n\left(c|w,I_t\right) > \exp(\text{-}k). \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The softmax loss in Equation 6 enables models to learn features and weights without prior observation of unlabeled samples. Such a function helps to curb the missing pseudo-label problem prevalent in conventional self-training and expectation maximization methods. To also prevent the situation where a model latches on to classes with large-samples, resulting in biased learning, the proposed approach introduces $k|\hat{\mathcal{Y}}_{t,n}|$. This factor determines the size of pseudo-labels to be selected from each class as well as assigning pseudo-labels to a sample. In Equation 6, the output probability $(p_n(c|w, I_t))$ must not be less than $\exp(\text{-}k)$, else it is assigned a zero-vector and ignored.

A vital component of the proposed method is the algorithm that determines the number of pseudo-labels to be added to the training data after each iteration (depicted in algorithm 2). The algorithm introduces k, that helps in determining the amount or rate pseudo-labeled samples to be selected to update the model of pseudo-labels as well as filtering out probabilities less than k. k is set by first taking the maximum probability on each sample, and these probabilities are then sorted across all samples and classes in a descending order. Then, k is set such that $\exp(-k)$ will be equivalent to the ranked probability

Algorithm 2: Algorithm for determining k in

```
input : Deep Learning Network $P(w)$, unlabeled
        Images $I_t$, selected pseudo-labels $p$
output: $k$
for $t \leftarrow 1$ to $T$ do
    $P_{I_t} = P(w, I_t)$;
    $M_{I_t} = \arg\max(P_{I_t}, axis = 0)$;
    $M = [M, matrix - to - vector(M_{I_t})]$
end
$M = sort(M, order = descending)$
$L = length(M) \times p$
$k = -\log(M[L])$ ;
$return(k)$
```

at $(p * T * N)$. p represents a portion number between $[0, 1]$. In this way, optimizing the pseudo-labels results in $p \times 100\%$ confident pseudo-labels to be used in training. The proposed selection algorithm allows the addition of the more pseudo-labels in the training sample for the next training iteration. M is the maximum probability output on each sample, and these probabilities are sorted across samples and classes.

IV. Materials and Experiments

A. Dataset

The dataset used in this work is obtained from [27]. It consists of 5,856 X-ray images. The images are anterior-posterior chest images that were taken chosen from retrospective pediatric patients between the ages of 1 and 5 years. The dataset ships with two kinds of chest x-ray images - normal and pneumonia stored in two separate folders. The number of normal images is significantly less than the number of the pneumonia (the normal class comprises only one-fourth of all data), creating a huge imbalance in the dataset. Sample images from the normal and pneumonia classes are depicted in Figure 2

B. Experimental Approach

This work proposes a CNN model that consists of five convolutional layers and one fully connected layer as the base learner for the self-training process. The CNN model is detailed as follows;

- First convolutional layer learns 64 filters, each of size 3 x 3
- Second convolutional layer learns 96 filters, each of size 3 x 3
- Third convolutional layer learns 128 filters, each of size 3 x 3
- Fourth convolutional layer learns 256 filters, each of size 3 x 3
- Fifth convolutional layer learns 256 filters, each of size 3 x 3

RELU activation is applied to every convolutional and fully connected layer. The RELU activation layer aids in faster convergence and also ensures that all negative activations are converted to zero. A batch normalization layer [28] is applied after every RELU activation layer.

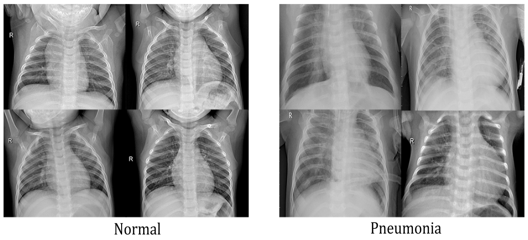

Fig. 2. Sample images from the normal and pneumonia classes

Batch normalization layers help normalize the activations of an input volume before passing activations to the next layer. Batch normalization layers are effective in reducing the number of epochs required to train a network, stabilizing the network, and also allow for a number of learning rates and regularization strengths. A pooling layer is applied after the batch normalization layer for the second and fifth convolutional layers. Pooling layers reduce the spatial size of the input volume, allowing for a reduction in the number of parameters. In the proposed architecture, max-pooling layers have a size of 2 x 2. Dropout with keep probability of 0.5 is applied after the fully connected layer.

$$\alpha = \textit{initLR} * (1 - \frac{epoch}{T_{epochs}})^p \tag{7}$$

initLR is the base learning rate, T_{epochs} is the total number of epochs, p is the exponential power, which is set to 1.

The network is trained from scratch and its weights are initializes using Gaussian distribution. The model is trained with the Adam optimizer [29] with a learning rate of 0.0001, $\beta_1 = 0.9$ and $\beta_2 = 0.99$. A polynomial decay learning rate scheduling is implemented since it allows for the decaying of the learning rate over a fixed number of epochs. The training process is for a total of 100 epochs with a batch size of 64. For data augmentation, random rotation with a range of 90°, and horizontal flipping have been implemented. Data augmentation helps curb overfitting in models. Input images are resized to 200 X 200 before being fed to the model.

For the training data, 70% is during training and 30% is reserved as test samples. The test samples are used as the unlabeled data for the self-training scheme. In all experiments, the CNN model is re-trained with hyper-parameters for top *k* using 5%, 10% and 20% of the pseudo-labeled samples of the unlabeled data. Experiments are performed using using Keras (version 2.2.4) [30] with Tensorflow backend (version 1.12) [31] and CUDA 9.0. The hardware platform for all experiments is an RTX 2080 graphic card with 8GB memory and a 32GB RAM. The overall workflow is illustrated in Figure 1

V. Results and Discussion

In this work, a self-paced learning scheme that integrates self-training for classifying anterior-posterior chest images as either normal or pneumonia was introduced. To augment the training data, the proposed approach utilized both labeled and unlabeled data in the training process. 30% of the dataset

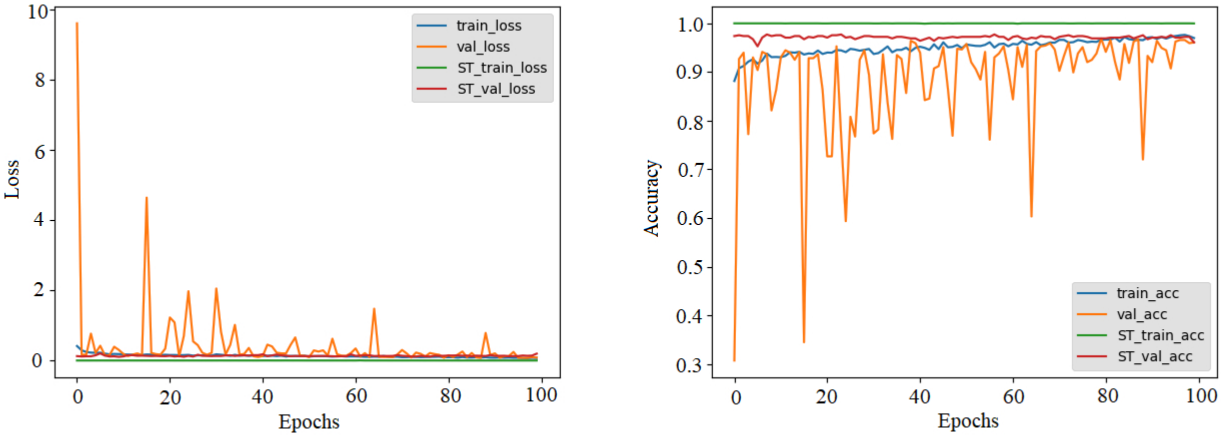

Fig. 3. Accuracy and loss plots after the training process. ST refers to the self-training plots. It is observed from the loss plot that, the loss value is initially high at the start of training but significantly decreases as training progresses. This high initial loss value is because, at the start of training, the model is only beginning to learn features or patterns from the data (since the model is being trained from scratch). By the end of 100th epoch, the loss is significantly lower. Observing the accuracy plots, it is observed that, both training and validation accuracy plots match up smoothly at the end of training, an indication that overfitting is effectively minimized.

TABLE I. Accuracy comparison of the proposed method with other works Compared to the supervised algorithms mentioned in the literature which uses only labeled data, the proposed approach shows significantly higher accuracy performance when only a portion of the generated pseudo-labels are used.

Method	Accuracy(%)
[27]	92.8
[32]	93.8
[10]	93.73
[19]	94.39
[18]	95.30
[17]	96.2
This work (Baseline, trained from scratch)	96.26
This work (All pseudo labels)	96.42
This work (top-5% pseudo-labels)	97.56
This work (top-10% pseudo-labels)	**98.04**
This work (top-20% pseudo-labels)	96.74

was reserved as the unlabeled data for the re-training process. For all experiments, the proposed approach was evaluated by using - i) using all generated pseudo-labels for the unlabeled data; and ii) using the top-5%, top-10% and top-20% confident pseudo-labels after setting a threshold k. Experimental results are shown in Table I. The best accuracy obtained was 98.04% when the top-10% most confident pseudo-labels were used. Using the top-5% confident pseudo-labels resulted in an accuracy of 97.56%, with the top-20% confident pseudo-labels yielding an accuracy of 96.74%. Using all the pseudo-labels yielded an accuracy of 96.42%. Training the baseline model from scratch resulted in an accuracy of 96.26%.

The unbalanced nature of the dataset is a challenge for deep learning models as such a scenario puts the model at the risk of overfitting on data. This is because, the model tends to be biased towards classes with more data representation. The proposed approach effectively curbs overfitting as shown by the accuracy and loss plots in Figure 3. The loss starts a high value because the model is trained from scratch and as such, at the initial training stage, the model is only getting to learn the data patterns. Over the course of the training process, there's a significant reduction in the loss value. The training and validation accuracy plots for both the baseline training and self-training indicate a near match-up of accuracies, an indication that overfitting is effectively minimized. The overall experimental results obtained demonstrate significant accuracy improvements though only a portion of the generated pseudo-labels were used, an indication of the strength of the proposed method.

A. *Comparison with Other Work*

A comparison of the proposed method with other methods mentioned in the literature is presented in this section. Table I shows the performance of the proposed approach in comparison with other works. It is pertinent to note that, the reported works in the literature adopt supervised learning techniques, where only labeled data is used in the training process, without the use of unlabeled data. The proposed method, which effectively and efficiently selects the most confident pseudo-labels as update to the model in the next training phase, outperforms the methods reported in the literature on the same dataset. In [17], the authors used a 16-layer pre-trained VGG model for classifying images as pneumonia or normal. A pre-trained model has been trained on the ImageNet dataset and such, it possesses a great deal of rich features. However, the proposed method yielded significantly higher results with a simple baseline model trained from scratch. Similarly, compared to works in the literature ([10],[18],[19]) that trained models on the same dataset with only labeled data, the proposed approach yields higher accuracy, rubber-stamping the point that, selecting the most confident pseudo-labels for training is pf significant contribution to the overall performance of a CNN learner.

VI. Conclusion

In this work, a self-paced learning scheme, which integrates self-training for classifying anterior-posterior chest images as either normal or pneumonia has been proposed. The proposed method utilizes both labeled and unlabeled data in the training process. A vital element of self-paced learning is that, it curbs the issue of mistake reinforcement learning, where a model incorrectly reinforces wrong predictions into a training set. As such, selecting the most confident pseudo-labels to augment the training set is a key step in ensuring the model generalizes well of data. To this end, this work proposed a novel pseudo-label generation and selection algorithm for selecting the top K most confident pseudo-labels to be added to the next training phase. Experiments with a simple CNN baseline model trained from scratch yielded significantly higher accuracies compared to other works mentioned in the literature, where only labeled data was used in the training process. Future work will seek to introduce more diversity into the self-paced learning process.

Acknowledgment

Sarpong Kwadwo Asare and Obed Tettey Nartey contributed equally to this work.

References

[1] World Health Organization, *Household Air Pollution and Health [Fact Sheet]*, WHO, Geneva, Switzerland, 2018.

[2] I. Rudan, L. Tomaskovic, C. Boschi-Pinto, and H. Campbell, “Global estimate of the incidence of clinical pneumonia among children under five years of age,” Bulletin of the World Health Organization, vol. 82, No. 12, Jan. 2004, pp. 895 - 903.

[3] S. Kiranyaz, T. Ince, and M. Gabbouj, ”Real-time patient-specific ECG classification by 1-D convolutional neural networks,” IEEE Trans. Biomed. Eng., vol. 63, No. 3, Mar. 2016, pp. 664 – 675.

[4] S. K. Asare, F. You, O.T. Nartey, “Efficient, Ultra-facile Breast Cancer Histopathological Images Classification Approach Utilizing Deep Learning Optimizers”, International Journal of Computer Applications, vol.177, No. 37, Feb. 2020, pp. 1 - 9.

[5] K. Sirinukunwattana, S. E. A. Raza, Y. W. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot, “Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images,” IEEE Trans. Med. Imaging, vol. 35, No. 5, May 2016, pp. 1196 - 1206.

[6] P. Huang, S. Park, R. Yan et al., “Added value of computer- aided CT image features for early lung cancer diagnosis with small pulmonary nodules: a matched case-control study,” Radiology, vol. 286, No. 1, Sept. 2017, pp. 286 – 295.

[7] A. R. Sadri, M. Zekri, S. Sadri, N. Gheissari, M. Mokhtari, and F. Kolahdouzan, “Segmentation of dermoscopy images using wavelet networks,” IEEE Trans. Biomed. Eng., vol. 60, no. 4, Apr. 2013, pp. 1134 – 1141. .

[8] Y. Yuan and Y.-C. Lo, ”Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks,” IEEE J. Biomed. Health Inform., vol. 23, no. 2, Mar. 2019, pp. 519 – 526.

[9] G. Litjens, T. Kooi, B.E. Bejnordi, S. Aaa, F. Ciompi, M. Ghafoorian, V.D.L. Jawm, G.B. Van, C.I. SÃ nchez, “A survey on deep learning in medical image analysis,” Med. Image Anal. vol. 42, Dec. 2017, pp. 60 – 88.

[10] S. Okeke, S. Mangal, J.M. Uchenna, and J. Do-Un, “An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare,” Journal of Healthcare Engineering, vol. 2019, March 2019, pp. 1 - 7.

[11] M. Kumar, B. Packer, and D. Koller, “Self-paced learning for latent variable models,” in Proc. Adv. Neural Inf. Process. Syst., Jan. 2010, pp. 1189 – 1197.

[12] U. Avni, H. Greenspan, E. Konen, M. Sharon, and J. Goldberger, “X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words,” Med Imaging, IEEE Transactions, vol. 30, No. 3, March 2011, pp. 733 - 746.

[13] J. Melendez, G. B. Van, P. Maduskar et al., “A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-ray,” IEEE Transactions on Medical Imaging, vol. 34, no. 1, 2015, pp. 179 – 192.

[14] S. Jaeger, A. Karargyris, S. Candemir et al., “Automatic tuberculosis screening using chest radio-graphs,” IEEE Transactions on Medical Imaging, vol. 33, no. 2, 2014, pp. 233 – 245.

[15] Z. Xue, D. You, S. Candemir et al., “Chest x-ray image view classification,” in Proceedings of the Computer-Based Medical Systems IEEE 28th International Symposium, São Paulo, Brazil, June 2015.

[16] P. Rajpurkar et al., “CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning,” ArXiv, vol. abs/1711.05225, Dec. 2017.

[17] S. Rajaraman, S. Candemir, I. Kim, G. Thoma, S. Antani, “Visualization and Interpretation of Convolutional Neural Network Predictions in Detecting Pneumonia in Pediatric Chest Radiographs,” Appl. Sci., vol. 8, 2018, pp. 1715.

[18] A.A. Saraiva et al., “Classification of Images of Childhood Pneumonia using Convolutional Neural Networks,” BIOIMAGING, Jan. 2019, pp. 112 - 119.

[19] S. Raheel, “Automated Pneumonia Diagnosis using a Customized Sequential Convolutional Neural Network,” ICDLT, Sept. 2019, pp. 64 - 70.

[20] J. Ureta, O. Aran, and J.P. Rivera, “Detecting pneumonia in chest radiographs using convolutional neural networks”, Proc. SPIE 11433, Twelfth International Conference on Machine Vision (ICMV 2019), Jan. 2020. pp. 116.

[21] X. Zhu, “Semi-supervised learning literature survey,” Ph.D. dissertation, Dept. Comput. Sci., Univ. Wisconsin, Madison, WI, USA, Tech. Rep. 07, vol. 2, 2008.

[22] F. Schwenker and E. Trentin, “Pattern classification and clustering: A review of partially supervised learning approaches,” Pattern Recognit. Lett., vol. 37, Feb. 2014, pp. 4 – 14, .

[23] O. T, Nartey, G. W. Yang, J. Z. Wu, and S. K. Asare,“Semi-Supervised Learning for Fine-Grained Classification With Self-Training,” *IEEE Access*, vol. 8, Jan. 2020, pp. 2109 - 2121.

[24] N. Gu, M. Fan, and D. Meng, ”Robust Semi-Supervised Classification for Noisy Labels Based on Self-Paced Learning” IEEE Signal Processing Letters, vol. 23, no. 12, Dec, 2016. pp. 1806 - 1810.

[25] E. Sangineto, M. Nabi, D. Culibrk, and N. Sebe, ”Self Paced Deep Learning for Weakly Supervised Object Detection”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 3, March, 2019, pp. 712 - 725.

[26] L. Jiang, D. Meng, S.I. Yu, Z. Lan, S, Shan, A.G. Hauptmann, ”Self-Paced Learning with Diversity”, In NIPS’14: Proceedings of the 27th International Conference on Neural Information Processing Systems, vol. 2, Dec. 2014, pp. 2078 – 2086.

[27] D. K. Kermany and M. Goldbaum, Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification, Mendeley Data, London, UK, 2018.

[28] S. Ioffe and C. Szegedy, ”Batch normalization: Accelerating deep network training by reducing internal covariate shift”, in 32nd International Conference on Machine Learning, ICML Feb. 2015.

[29] D. P. Kingma and J. L. Ba, ”Adam: a Method for Stochastic Optimization,” Int. Conf. Learn. Represent. 2015.

[30] F. Chollet, ”Keras: Deep Learning for humans,” Github, 2015.

[31] M. Abadi et al., ”TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems,” 2016.

[32] S. S. Yadav, S. M. Jadhav, “Deep convolutional neural network based medical image classification for disease diagnosis”, Journal of Big Data, vol. 6, Dec. 2019, pp. 113.