IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 May 2023, accepted 18 June 2023, date of publication 26 June 2023, date of current version 3 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3289219

RESEARCH ARTICLE

DL4ALL: Multi-Task Cross-Dataset Transfer Learning for Acute Lymphoblastic Leukemia Detection

ANGELO GENOVESE[1], (Senior Member, IEEE), VINCENZO PIURI[1], (Fellow, IEEE), KONSTANTINOS N. PLATANIOTIS[2], (Fellow, IEEE), AND FABIO SCOTTI[1], (Senior Member, IEEE)

[1]Department of Computer Science, Università degli Studi di Milano, 20133 Milan, Italy
[2]Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada

Corresponding author: Angelo Genovese (angelo.genovese@unimi.it)

This work was supported in part by the European Commission (EC) through EdgeAI Project under Grant 101097300 and through GLACIATION Project under Grant 101070141, and in part by the Italian Ministero dell'Università e della Ricerca (MUR) through SERICS Project [National Recovery and Resilience Plan (NRRP) MUR Program by the EU—Next Generation EU (NGEU)] under Grant PE00000014.

ABSTRACT Methods for the detection of Acute Lymphoblastic (or Lymphocytic) Leukemia (ALL) are increasingly considering Deep Learning (DL) due to its high accuracy in several fields, including medical imaging. In most cases, such methods use transfer learning techniques to compensate for the limited availability of labeled data. However, current methods for ALL detection use traditional transfer learning, which requires the models to be fully trained on the source domain, then fine-tuned on the target domain, with the drawback of possibly overfitting the source domain and reducing the generalization capability on the target domain. To overcome this drawback and increase the classification accuracy that can be obtained using transfer learning, in this paper we propose our method named "Deep Learning for Acute Lymphoblastic Leukemia" (DL4ALL), a novel multi-task learning DL model for ALL detection, trained using a cross-dataset transfer learning approach. The method adapts an existing model into a multi-task classification problem, then trains it using transfer learning procedures that consider both source and target databases at the same time, interleaving batches from the two domains even when they are significantly different. The proposed DL4ALL represents the first work in the literature using a multi-task cross-dataset transfer learning procedure for ALL detection. Results on a publicly-available ALL database confirm the validity of our approach, which achieves a higher accuracy in detecting ALL with respect to existing methods, even when not using manual labels for the source domain.

INDEX TERMS Acute lymphoblastic leukemia (ALL), deep learning (DL), convolutional neural networks (CNNs).

I. INTRODUCTION

Acute Lymphoblastic (or Lymphocytic) Leukemia (ALL) is a disease which affects the blood cells, can spread rapidly in the body, and may result in a fatal outcome if left undiagnosed and untreated. It is therefore important to detect the presence of ALL as soon as possible, in particular one of the main steps in detecting its presence is the analysis of White Blood Cells (WBC) in peripheral blood samples. Such analysis, usually performed manually by an expert pathologist, has the purpose of detecting the presence of lymphoblasts, which are WBCs with malformations. Although lymphoblasts occur normally in the bone marrow, a greater concentration of lymphoblasts in peripheral blood, with respect to standard levels, can be associated with the presence of ALL [1], [2], [3]

The main problem with manually analyzing WBCs is that the process is time consuming and repetitive, therefore easily

The associate editor coordinating the review of this manuscript and approving it for publication was Chuan Li.

causing fatigue and a decreasing accuracy in labeling samples when more time is spent in the process. Hence, to partially automate the inspection process, an increasing number of methods in the literature are considering Computer Aided Diagnosis (CAD) systems, which often use techniques such as image processing and Machine Learning (ML) to automatically classify WBCs [4], [5]. Within ML-based CADs, the majority of the methods are considering Deep Learning (DL) and Convolutional Neural Networks (CNN), because of their high accuracy in several fields and their capability of automatically learning data representations, without the need for handcrafted feature extraction [6], [7].

CAD systems based on DL for the detection of ALL are being proposed in the literature for increasing the accuracy and reliability of the classification [5], [8], [9], by introducing original learning procedures [2], [10], [11], [12], ad-hoc network architectures [13], [14], [15], or DL-based preprocessing [1]. Approaches that introduce original learning procedures usually consider transfer learning [1], [2], [10], [16], due to the limited dimensionality of the ALL databases, as happens in several medical fields because of the scarcity of labeled samples [17], [18]. In particular, the histopathological transfer learning approach, based on pre-training the CNN on a histopathology database (*source domain*) and fine tuning it on the ALL database (*target domain*), achieved a state-of-the-art accuracy in the detection of ALL, because of the higher similarity of the source and target domains, with respect to using models pre-trained on the ImageNet database [2], [19]. Moreover, pre-training the CNN on different but related databases with respect to the one used for ALL classification has been associated with a greater generalization capability in medical imaging [17], [20], [21]. However, such pre-training uses a traditional transfer learning procedure, which requires the models to be fully trained on the source domain, then fine tuned on the target domain, and has the drawback that the CNN may not be able to fully adapt its structure on the target domain during the fine tuning phase, possibly overfitting the source domain and reducing the generalization capability on the target domain [22], [23].

To overcome the drawback of different source and target databases in transfer learning approaches for ALL detection, in this paper we introduce our method named ''Deep Learning for Acute Lymphoblastic Leukemia'' (DL4ALL),[1] a multi-task model with two separate fully-connected (FC) layers as outputs, one for the source domain and one for the target domain. Differently from the approaches in the literature for ALL detection, DL4ALL is trained by introducing three novel cross-dataset transfer learning procedures, namely *regular*, *greedy*, and *self-supervised*, which differ based on how the source domain and the corresponding labels are used. All procedures use both source and target domains at the same time, interleaving data batches during training even when the respective databases are significantly different from each other.

[1] https://iebil.di.unimi.it/cnnALL/index.htm

This work represents the first method in the literature that uses a multi-task cross-dataset transfer learning procedure for ALL detection. While methods based on multi-task cross-dataset learning have already been proposed in the literature [15], [20], [22], [23], [24], [25], such approaches do not consider the problem of ALL detection.

To evaluate the validity of the approach, we consider a histopathology database as a source domain and the ALL database as target domain. We chose the histopathology database given the high accuracy demonstrated in transfer learning for medical imaging [17] and for ALL detection [2], together with the fact that histopathological tissue labeling and cancer detection are two interrelated problems [26]. In training, the model can therefore take advantage of both databases, without a pre-training step on the source database which could excessively bias the model towards the source domain. In testing, we apply the trained model on the ALL database, with the purpose of classifying each WBC sample as either ''normal'' or ''lymphoblast''. We evaluate our method using recent databases for histopathology tissue type classification and ALL classification, two different DL models (CNN and attention-based), and three different cross-dataset transfer learning procedures (*regular*, *greedy*, and *self-supervised*), obtaining superior results in ALL detection with respect to the state-of-the art.

The remainder of the paper is structured as follows. Section II reviews the related works. Section III introduces the methodology. Section IV describes the experimental results. Finally, Section V concludes the work.

II. RELATED WORKS

In this section we first review the most recent approaches for ALL detection and then present an overview of the techniques for learning with limited labeled data.

A. ACUTE LYMPHOBLASTIC LEUKEMIA DETECTION

When considering ML approaches for CAD systems and ALL detection, traditional approaches usually describe a handcrafted feature extraction step. However, recent approaches for medical imaging and ALL detection have been almost exclusively focusing on DL, which in most cases is able to learn representations directly from data [27] and does not require handcrafted features [5], [28]. Therefore, in this paper we will review only the methods using DL-based models. In particular, it is possible to divide DL-based methods for ALL detection in three categories, based on the approach used to achieve a more accurate classification of WBCs [1], [2]: *i)* original learning procedures; *ii)* ad-hoc network architectures; *iii)* DL-based preprocessing.

The approaches belonging to *i)* include methods that pre-train a CNN on databases containing general purpose images (e.g., ImageNet) and then fine tune it on the target ALL dataset, such as the works described in [10], [11], [16], and [28]. A similar method is described in [12], with the difference that, after the pre-training step, the method applies swarm optimization to perform a feature selection

that better adapts the CNN to classify ALL samples. The works described in [2], [19], and [29] also perform a pre-training of the CNN but on histopathology images, achieving a more accurate classification than methods performing a pre-training using general purpose images.

The methods belonging to *ii)* introduce ad-hoc modifications of existing CNN architectures, with the purpose of better adapting them for the classification of ALL. For example, the approach proposed in [15] proposes a modification of the ResNet architecture [30] that is able to extract features from WBC samples at both the local and global level. Other approaches describe ad-hoc architectures to reduce overfitting in the case of small datasets, for example considering Bayesian CNNs [31] or shallow CNNs with a reduced number of layers with respect to a ResNet [32].

The approaches belonging to *iii)* describe methods that use DL to preprocess the images with the aim of enhancing the details of the WBC samples, such as the methodology introduced in [13], which describes a convolutional layer specifically designed to perform a stain deconvolution and normalize the colors. Differently, the approach presented in [1] uses a procedure based on CNN to perform an adaptive and intelligent tuning of the unsharpening algorithm, to normalize the focus quality of ALL images.

B. LEARNING WITH LIMITED LABELS

Training DL using supervised learning techniques enables, in the majority of situations, to obtain the best classification accuracy on the considered databases. For example, the accuracy on the ImageNet database has greatly improved by using deeper DL models such as CNNs and attention-based mechanisms in a supervised learning fashion [33]. However, supervised deep learning requires the labeling (often manual) of an extensive number of samples, using an expensive and time consuming processing. In many application scenarios (e.g., medical imaging), there is a limited availability of such labeled samples, with the number of available labels greatly inferior to general purpose databases such as ImageNet.

To overcome the scarcity of labeled data, several approaches have been proposed in the literature, including data augmentation, domain adaptation, few-shot learning, multi-task learning, semi/weakly-supervised learning, unsupervised learning, and self-supervised learning [33], [34]. Among the above mentioned techniques, data augmentation and domain adaptation using supervised fine tuning (often referred to as ''transfer learning'') are considered as common practice when designing DL-based methods, for example considering a CNN pre-trained on a source domain (e.g., ImageNet) and fine tuned on a target domain, with the images randomly rotated, flipped, or cropped during the tuning process [35]. Other supervised domain adaptation methods include knowledge distillation (e.g., the teacher-student method), in which the knowledge from a larger model is ''distilled'' into a smaller one to avoid overfitting and reducing the computational complexity when training on the target database [36], [37].

When very few labeled samples are available for the target domain, few-shot learning techniques have been proposed to use as much as possible the knowledge of the pre-trained model to generalize to unseen data [36], [38]. Another approach that has been proposed in the case of few labeled samples is multi-task learning, which consists in training the model on multiple tasks at the same, for example by using multiple datasets with limited samples, to force the model to learn a general representation and limit overfitting [15], [20].

In the cases where the target data contains both labeled and unlabeled samples, some approaches consider semi/weakly-supervised learning, which can also leverage unlabeled samples when performing a domain adaptation, for example analyzing them in the latent space and automatically assigning them a label based on the closest sample [39].

When only unlabeled samples are available, unsupervised learning methods can be used to extract knowledge and perform decisions based on the underlying structures within the data. Examples of traditional unsupervised learning include approaches for dimensionality reduction (e.g., PCA) and data clustering (e.g., self-organizing maps), while recent methods include DL-based approaches for replicating complex data distribution and reduce noise (e.g., autoencoders, generative adversarial networks) [34]. Unsupervised learning techniques are often useful when performing domain adaptation using DL models: since the objective is to accurately classify samples in the target domain, the need for labels in the source domain is reduced. For example, it is possible to train an autoencoder to replicate unlabeled data from the source domain, then fine tune it on the labeled data of the target domain in a supervised way [40]. Combining supervised and unsupervised approaches, self-supervised learning methods are being increasingly considered due to their advantage of performing a supervised learning but using pseudo labels, which can also be automatically generated, without the need for a manual labeling process [41].

In this paper, we propose an innovative method for ALL detection that considers recent advances in DL approaches, such as multi-task learning and self-supervised learning, to cope with the limited availability of labeled samples in ALL database. To the best of our knowledge, this is the first method in the literature for ALL detection considering a multi-task architecture trained using a cross-dataset transfer learning procedure that uses both source and target domains at the same time.

III. METHODOLOGY

This section describes the proposed methodology for ALL detection based on DL4ALL, consisting of a multi-task model trained using three multi-task cross-dataset transfer learning procedures. Our method considers a model with an existing architecture, then creates a multi-task learning architecture by substituting the last FC layer with two layers, respectively one for the source domain and one for the target domain.

The learning phase is then performed using a cross-dataset transfer learning procedure that uses source and target domains at the same time, in which the batches for each database are extracted in an interleaved fashion to compute the loss and adjust the weights of the model.

We propose three novel procedures for ALL detection, *regular*, *greedy*, and *self-supervised*, which differ in the way they use the data from the source domain, by considering different amounts of source data, different labels, and different levels of supervised learning. Lastly, the trained model is applied only on the target domain to perform the ALL detection, by classifying each WBCs sample as either ''normal'' or ''lymphoblast''.

Our method executes the following steps: *A)* creation of DL4ALL; *B)* cross-dataset transfer learning; *C)* deep ALL classification. Fig. 1 shows the outline of the methodology.

A. CREATION OF DL4ALL

To make the proposed method applicable to any DL-based model, we create the DL4ALL by starting from an existing deep architecture (e.g., CNN-based [33], attention-based [42]). In this way, it is also possible to consider pre-trained architectures (e.g., models pre-trained on ImageNet are widely available [43]). To create the DL4ALL, as a first step we remove the last fully connected layer of the chosen model, which usually has a number of neurons equal to the number of possible classes. As an example, in most cases the last FC layer has an output size of 1000, corresponding to the 1000 classes of the ImageNet database.

As a second step, we create the novel multi-task learning architecture for ALL detection by connecting two FC layers in parallel, one for the source domain and one for the target domain. Each FC layer is then responsible for classifying samples of the database in the corresponding domain. As shown in Fig. 2, one FC layer outputs the classification of the histopathological tissue type (*source domain*, shown in blue in the figure) and one FC layer outputs the classification of the ALL (*target domain*, shown in green in the figure).

Lastly, we apply a sigmoid layer after the FC layer corresponding to the histopathological database, since such database has samples with multiple labels [44]. Differently, we apply a softmax layer after the FC layer corresponding to the ALL database, since such database has samples each belonging to a single class [45].

B. CROSS-DATASET TRANSFER LEARNING

To train the DL4ALL, we propose three innovative cross-dataset transfer learning procedures for ALL detection, all of which use both source and target domains at the same time, considering batches from the two databases in an interleaved fashion. In this way, we take advantage of both databases without the model biasing on the source domain.

While traditional transfer learning procedures perform a full training on the source domain followed by a deep tuning on the target domain, our method uses a training procedure –based on gradient descent– in which the odd-numbered batches are pulled from the source domain and the even-numbered batches are pulled from the target domain. As mentioned in the introduction, we consider a histopathology database as a source domain and the ALL database as target domain, given that transfer learning procedures with histopathology databases as source domain are proven to increase classification accuracy in the case of ALL detection [2], [19], [26].

We propose three novel cross-dataset transfer learning procedures for ALL detection, which differ in the way the source domain is used to help increasing the classification accuracy in the target domain, by considering different amounts of source data, different labels, and different levels of supervised learning:

- *Regular* cross-dataset transfer learning;
- *Greedy* cross-dataset transfer learning;
- *Self-supervised* cross-dataset transfer learning.

1) REGULAR CROSS-DATASET TRANSFER LEARNING

The first procedure we propose consists in training the DL4ALL by interleaving batches from databases of the source and the target domains, respectively, computing the loss for each batch, and considering an aggregated global loss as the weighted sum of the two losses. For each epoch, the procedure is based on the following steps:

1) *Forward pass (source domain)*: we extract a batch from the source database, apply the DL4ALL, and consider the output of the corresponding FC layer.
2) *Loss computation (source domain)*: based on the output of the FC layer corresponding to the source domain, we compute the multi-label loss L_{source} following Eqn. 1:

$$\begin{aligned} &L_{source}(x, y) \\ &= -\frac{1}{C}\sum_{i} w[i]\, y[i] \log\left((1+\exp(-x[i]))^{-1}\right) \\ &\quad + (1-y[i]) \log\left(\frac{\exp(-x[i])}{(1+\exp(-x[i]))}\right), \end{aligned} \tag{1}$$

where the choice of using a multi-label loss is caused by the histopathological database in the source domain having samples with multiple possible labels each [44].
3) *Batch normalization update*: we extract a batch from the target database and we update the parameters of the batch normalization layer by performing a preliminary forward pass without computing the loss. In fact, the parameters of the batch normalization layer are updated during the forward pass and not during backpropagation. Without a preliminary forward pass, the model would compute the output of the forward pass in the target domain using the batch normalization parameters tuned for the source domain, resulting in lower accuracies [20].

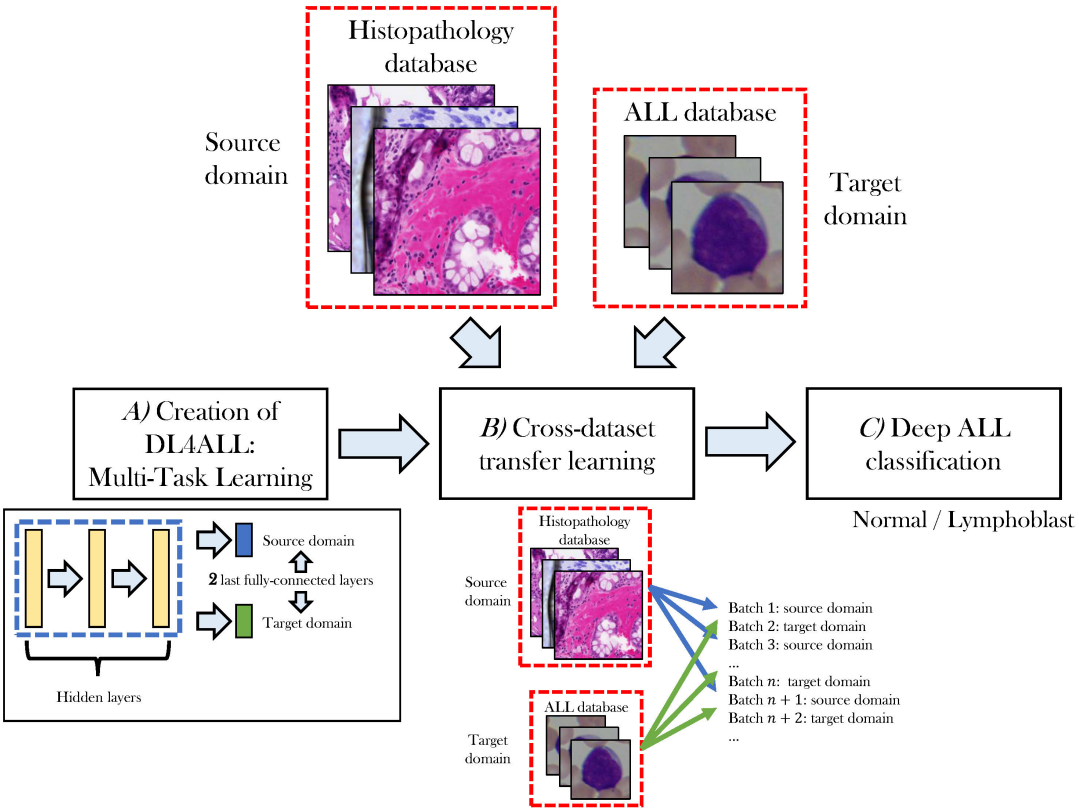


FIGURE 1. Outline of the proposed methodology. The method executes the following steps: *A)* creation of DL4ALL, in which we adapt the model to perform a multi-task learning; *B)* cross-dataset transfer learning, in which we interleave data from source and target domains during the training; *C)* deep ALL classification, in which we apply the trained model on the ALL images to predict the presence of a lymphoblast.

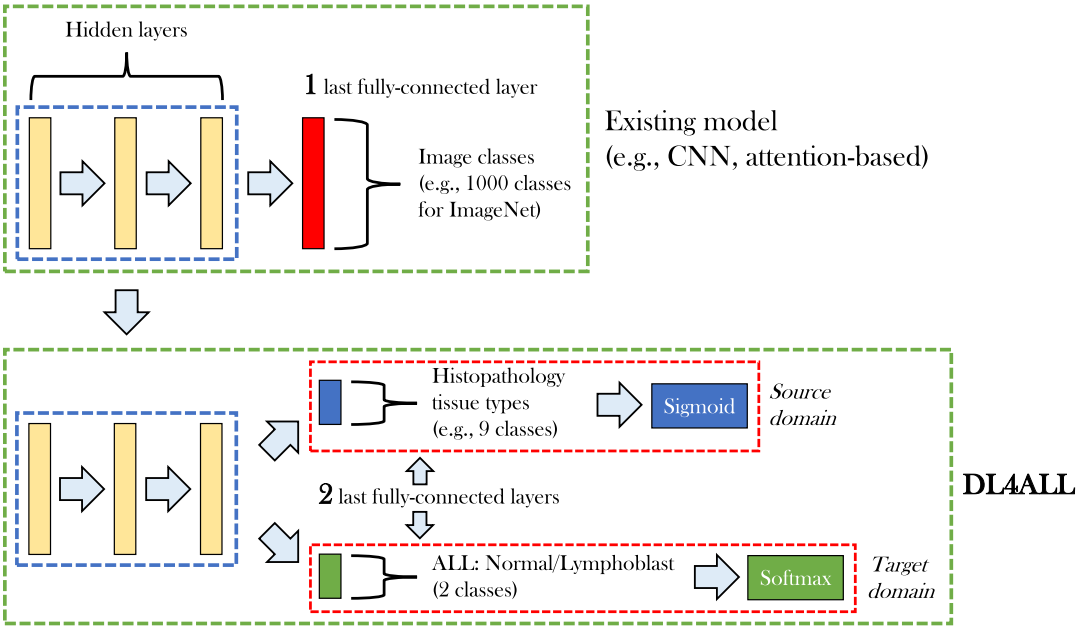


FIGURE 2. DL4ALL: starting from an existing deep architecture, first we remove the last fully connected layer of the chosen model. Second, we create the novel multi-task learning architecture by connecting two fully-connected layers (FC) in parallel.

4) *Forward pass (target domain)*: we extract a batch from the target database, apply the DL4ALL, and consider the output of the corresponding FC layer.

5) *Loss computation (target domain)*: based on the output of the FC layer corresponding to the target domain, we compute the cross-entropy loss L_{target} as described

in Eqn. 2:

$$L_{target}(x, class) = -w[i] \log\left(\frac{\exp(x[class])}{\sum_j \exp(x[j])}\right), \quad (2)$$

where the choice of using a cross-entropy loss is caused by the ALL database in the target domain having samples that each belong to a single class.

6) *Aggregated global loss*: we compute the aggregated global loss as the weighted sum of L_{source} and L_{target}, according to Eqn. 3:

$$L_{global} = w_1 \cdot L_{source} + w_2 \cdot L_{target}, \quad (3)$$

with w_1, w_2 chosen experimentally.

7) *Gradient normalization*: following the stochastic gradient descent (SGD) algorithm, we compute the gradients based on the L_{global} loss. Then, we take into account that the weights of the last two FC layers of DL4ALL may see a different number of samples, since the batch size may be different in the source and target databases [20]. Therefore, we normalize the gradients of the FC layers by considering the respective batch sizes, following Eqn. 4:

$$\partial_{FC,source} = \partial_{FC,source} \cdot \frac{|B_{source}| + |B_{target}|}{|B_{source}|};$$
$$\partial_{FC,target} = \partial_{FC,target} \cdot \frac{|B_{source}| + |B_{target}|}{|B_{target}|}, \quad (4)$$

where $\partial_{FC,source}$ and $\partial_{FC,target}$ indicate the gradients of the FC layers for the source and target domain, respectively, while $|B_{source}|$ and $|B_{target}|$ describe the cardinality of the batches for the source and target databases, respectively.

8) *Backpropagation*: following the SGD algorithm, we update the weights of DL4ALL considering the computed gradients.

The pseucodode for the *regular* cross-dataset transfer learning procedure is described in Alg. 1. The steps $1-8$ consist in a single iteration of the algorithm. An epoch of the training algorithm iterates the algorithm until there are no more batches to extract. After each epoch, we validate the model on the validation subset of the target domain (ALL database). After the last epoch, we keep the weights of the model for which we achieved the greatest validation accuracy.

2) GREEDY CROSS-DATASET TRANSFER LEARNING

The second procedure we propose consists in training the DL4ALL by also considering both source and target domains at the same time and interleaving batches. However, differently than the *regular* cross-dataset transfer learning (Section III-B1), at each iteration we consider each batch from the source domain only if it helps reduce the target loss L_{target}, following the greedy design paradigm [46].

For each iteration of the algorithm, after the forward pass in the source domain, we compute the L_{source} and perform the backpropagation, obtaining a DL4ALL whose weights have been updated only considering the source domain. Then, we update the batch normalization parameters, perform the forward pass in the target domain, and compute L_{target}. If L_{target} is reduced with respect to the previous iteration of the algorithm, we compute L_{global}, normalize the gradients, and perform the backpropagation, obtaining a DL4ALL with weights updated considering both domains, similarly to the *regular* cross-dataset transfer learning.

The main difference with respect to the *regular* cross-dataset transfer learning is that if L_{target} is not reduced with respect to the previous iteration, we roll back DL4ALL to the state it was before backpropagating L_{source}, and only consider the loss of the target domain when backpropagating $L_{global} = L_{target}$. The outline of the *greedy* cross-dataset transfer learning procedure is shown in Fig. 3.

3) SELF-SUPERVISED CROSS-DATASET TRANSFER LEARNING

The third procedure we propose consists in training the DL4ALL using a cross-dataset transfer learning with similar steps as the *regular* procedure described in Section III-B1. However, differently from the *regular* cross-dataset transfer learning, in this procedure we consider a *self-supervised* approach, rather than a supervised learning approach, to compute L_{source}, without the need to use manually-obtained labels at the sample level for the source domain.

The *regular* cross-dataset transfer learning procedure takes advantage of each sample in the source domain being manually labeled using fine-grained annotations. Such labeling enables to consider supervised learning approaches and obtain a high classification accuracy, since supervised learning usually performs better than unsupervised learning [47]. When a histopathological database is considered in the source domain, as it is in this paper, it is then possible to obtain a high classification accuracy of histological tissues. However, the labeling is extremely time-consuming and expensive to obtain, since expert personnel is required to perform a manual classification of histological tissues [44].

In the case of transfer learning, when using the histopathological database as a source domain, the need for long and expensive labeling is reduced, since the interest is towards achieving an increased accuracy in the target domain. Therefore, to avoid the need to have manually-annotated labels for each sample in the source domain, we propose a *self-supervised* cross-dataset transfer learning procedure, where we compute L_{source} without using the manual labels, but instead by considering self-supervised labels extracted from the data itself.

To compute the self-supervised labels, we consider as source domain a histopathological database composed of several image patches, obtained by cropping whole slide images (WSI) into smaller areas. For example, in our work we consider a database where 100 WSIs are divided into 17, 668 patches. Since current deep models are not able to directly process WSIs due to their extremely high dimensions, the models need to process each patch separately.

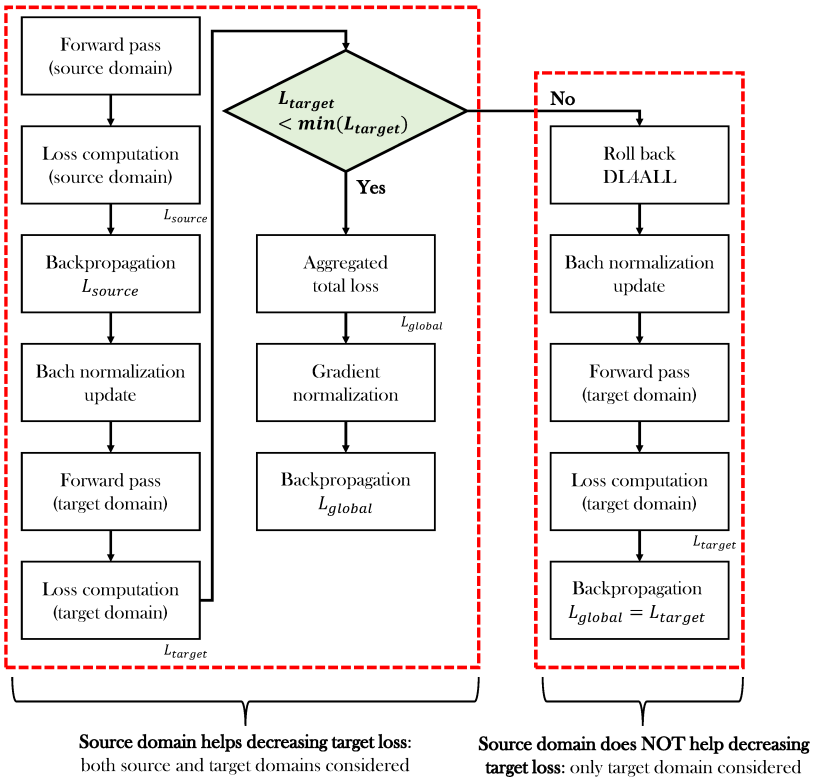


FIGURE 3. Outline of a single iteration of the proposed *greedy* cross-dataset transfer learning procedure: after the forward pass in the source domain, we backpropagate the loss of the source domain and compute the target loss. If the resulting target loss is reduced with respect to the previous iteration of the algorithm after considering the source domain (*source domain helps decreasing target loss*), we keep the model and backpropagate the target loss as well, otherwise (*source domain does not help decreasing target loss*) we roll back DL4ALL and consider only the target domain.

To use supervised learning approaches, each patch has to be manually labeled [44]. However, instead of using patch-level labels, we obtain self-supervised labels by considering WSI-level labels. In particular, for each patch we consider as label the progressive number indicating which WSI it was cropped from. Such number can be easily extracted by the filename of each patch. Moreover, it is a label that can be applied with limited effort and even by non-expert personnel.

The difference with the *regular* procedure described in Section III-B1 is in Step 2:

2) *Loss computation (source domain)*: based on the output of the FC layer corresponding to the source domain, we compute the cross-entropy loss L_{source}, following Eqn. 5:

$$L_{source}(x, class) = -\log\left(\frac{\exp(x[class])}{\sum_j \exp(x[j])}\right), \quad (5)$$

where *class* is the progressive number indicating which WSI each image patch was cropped from.

C. DEEP ALL CLASSIFICATION

To perform the final deep ALL classification, we apply the DL4ALL model –trained using the procedures described in Section III-B– on the testing subset of the target domain. We obtain three different DL4ALL, one for each procedure proposed: DL4ALL $_{reg}$, DL4ALL $_{greedy}$, and DL4ALL $_{self}$. Each DL4ALL model gives two outputs (output$_{source}$ and output$_{target}$), one for each of the two FC layers. However, we are interested only in classifying samples in the target domain, so we discard output$_{source}$ and compute the error measures considering output$_{target}$. The output for each image in the target domain (ALL) is then a binary number that indicates the predicted presence of a lymphoblast (0: *normal*; 1: *lymphoblast*).

IV. EXPERIMENTAL RESULTS

This section describes the experimental results, including the used databases, the model and training parameters, the error measures, and the results both in terms of quantitative and qualitative evaluation.

A. USED DATABASES

In this work we consider two databases, one for the source and one for the target domain respectively. For the source domain, we use a histopathology database, the Atlas of Digital

Algorithm 1 Pseudocode Describing the Proposed *regular* Cross-Dataset Transfer Learning Procedure

```
model = initializeWeigths(model);
while epoch < maxEpochs do
    while batches < numBatches do
        // 1. Forward pass (source domain)
        batch_source = load(sourceDomain, batchSize_source);
        input, target = decomposeBatch(batch_source);
        output_source, output_target = model(input);
        // 2. Loss computation (source domain)
        L_source = MultiLabelLoss(output_source, target);
        // 3. Batch normalization update
        batch_target = load(targetDomain, batchSize_target);
        input, target = decomposeBatch(batch_target);
        output_source, output_target = model(input);
        // 4. Forward pass (target domain)
        batch_target = load(targetDomain, batchSize_target);
        input, target = decomposeBatch(batch_target);
        output_source, output_target = model(input);
        // 5. Loss computation (target domain)
        L_target = CrossEntropyLoss(output_target, target);
        // 6. Aggregated total loss
        L_global = w1 · L_source + w2 · L_target;
        // 7. Gradient normalization
        ∂, ∂_FC,source, ∂_FC,target = computeGradients(L_global, DL4ALL);
        ∂_FC,source = normGradients(batchSize_source);
        ∂_FC,target = normGradients(batchSize_target);
        // 8. Backpropagation
        DL4ALL = updateWeights(∂, ∂_FC,source, ∂_FC,target);
    end
end
```

Pathology (ADP) [44],[2] containing 17,668 RGB image patches $\{p\}$ extracted from 100 WSIs, with image size 272 × 272 pixels. Each patch p has been manually labeled according to three levels of labeling, with each level having a more detailed classification. Therefore, each patch p is associated to a set of labels $L(p) = \{l_1, l_2, l_3\}$, where l_1 corresponds to the most coarse classification, l_2 to the intermediate classification, and l_3 to the most detailed classification. Within each level, since each patch can describe multiple histological tissues, the labels are not mutually exclusive, and each patch can be associated to multiple labels [44]. Then, each level of classification has a different number of output classes, for example the first level of labels l_1 has 9 possible outputs, hence in l_1 is described by a vector with 9 elements $|\{l_1\}| = 9$. Moreover, each patch is also associated with the self-supervised label l_{self}, indicating the progressive number of the WSI it was extracted from, with $l_{self} \in [1, 100]$. Fig. 4 shows examples of patches from the ADP database and the corresponding labels, while Table 1 lists the class distribution of samples for the database.

For the target domain, we use an ALL database, the C_NMC_2019 Dataset from the ALL Challenge in ISBI 2019 [13],[3] containing 10,661 RGB samples of WBCs, with image size 450 × 450 pixels, divided in two classes (0: *normal*; 1: *lymphoblast*). The images have been cropped to show only the region of interest surrounding the cell. Table 2 lists the class distribution of samples for the database.

B. MODEL AND TRAINING PARAMETERS

In this paper, we consider two models, a ResNet18 CNN [30] and the Vision Transformer (ViT), an attention-based model for image classification [42]. We chose the ResNet18 since it is one of most used CNNs and it exhibits high accuracy for histopathological image classification and ALL detection, also in transfer learning configurations [2], [26], [44]. Moreover, we considered the ViT since it exhibited state-of-the-art performance for image classification and object recognition, especially in transfer learning configurations, when pretrained on large databases such as the ImageNet-21k [43]. In particular, we consider a ViT which divides the input image in patches of 16 × 16 pixels, with 12 heads, 12 layers, hidden size = 768, and MLP size = 3072. We considered the default parameters of the model. Both the ResNet18 and ViT are pretrained on the ImageNet database [48].

We split both the ADP and the CNMC databases using 70% data for training, 20% for validation, and 10% for testing. Before training, we normalize data to have 0 mean and 1 standard deviation, with normalization parameters computed on the training subsets of the respective databases. Then, we perform data augmentation on the training subsets, by randomly applying rotations, horizontal flips, and vertical flips. We repeat the training and testing 5 times and average the results.

After creating the multi-task DL4ALL using the procedure described in Section III-A, we apply a warmup phase in which we train only the last 2 FC layers of DL4ALL separately using a standard SGD. In particular, first we apply the SGD for 1 epoch to the FC layer corresponding to the source domain (the remaining layers are frozen). Second, we apply the SGD for 1 epoch to the FC layer corresponding to the target domain (the remaining layers are frozen). Such warmup phase is used to reduce the problem of large gradients initially flowing

[2] https://www.dsp.utoronto.ca/projects/ADP

[3] https://wiki.cancerimagingarchive.net/display/Public/C_NMC_2019+Dataset%3A+ALL+Challenge+dataset+of+ISBI+2019

TABLE 1. Number of samples for each class in the histopathology database Atlas of Digital Pathology (ADP), according to each level of labeling l_1, l_2, l_3.

Class	Level of labeling			Num. of samples
	l_1	l_2	l_3	
1	Epithelial (E)	Simple Epithelial (E.M)	Simple Squamous Epithelial (E.M.S)	3341
2			Simple Cuboidal Epithelial (E.M.U)	5240
3			Simple Columnar Epithelial (E.M.O)	2533
4		Stratified Epithelial (E.T)	Stratified Squamous Epithelial (E.T.S)	355
5			Stratified Cuboidal Epithelial (E.T.U)	3662
6			Stratified Columnar Epithelial (E.T.O)	783
7			Stratified Epithelial Undifferentiated (E.T.X)	22
8		Pseudostratified Epithelial (E.P)		50
9	Connective Proper (C)	Dense Connective (C.D)	Dense Irregular Connective (C.D.I)	4481
10			Dense Regular Connective (C.D.R)	68
11		Loose Connective (C.L)		8768
12		Connective Proper Undifferentiated (C.X)		291
13	Blood (H)	Erythrocytes (H.E)		7504
14		Leukocytes (H.K)		1739
15		Lymphocytes (H.Y)		5232
16		Blood Undifferentiated (H.X)		126
17	Skeletal (S)	Mature Bone (S.M)	Compact Bone (S.M.C)	298
18			Spongy Bone (S.M.S)	233
19		Endochondral Bone (S.E)		38
20		Cartilage (S.C)	Hyaline Cartilage (S.C.H)	10
21			Cartilage Undifferentiated (S.C.X)	35
22		Marrow (S.R)		157
23	Adipose (A)	White Adipose (A.W)		536
24		Brown Adipose (A.B)		2
25		Marrow Adipose (A.M)		137
26	Muscular (M)	Smooth Muscle (M.M)		4213
27		Skeletal Muscle (M.K)		783
28	Nervous (N)	Neuropil (N.P)		2198
29		Neurons (N.R)	Nerve Cell Bodies (N.R.B)	1840
30			Nerve Axons (N.R.A)	59
31		Neuroglial Cells (N.G)	Microglial Cells (N.G.M)	593
32			Schwann Cells (N.G.W)	22
33			Neuroglial Cells Undifferentiated (N.G.X)	1856
34	Glandular (G)	Exocrine Gland (G.O)		6976
35		Endocrine Gland (G.N)		1115
36		Gland Undifferentiated (G.X)		66
37	Transport Vessel (T)			6045
-	Total			17668*

Notes. * The labels are not mutually exclusive.

TABLE 2. Number of samples for each class in the ALL database C_NMC_2019.

Class	Label	Num. of samples
0	Normal	7272
1	Lymphoblast	3389
-	Total	10,661

towards the last FC layers, since the convolutional layers might be pre-trained and are followed by FC layers with random initialization [20].

After the warmup phase, we train the models using the proposed cross-dataset transfer learning procedures described in Section III-B. We use a SGD for 100 epochs, with momentum $m = 0.9$, weight decay $5e^{-4}$, and batch size 8. We chose the number of epochs, momentum, and weight decay following common practices regarding learning with SGD [49], while we chose the batch size considering the amount of RAM in our GPUs. We consider different learning rates for the last FC layers and for the rest of the model. In particular, we consider $lr_{FC} = 1e^{-3}$ for the last FC layers and $lr_{shared} = 2e^{-4}$ for the rest of the model shared between tasks. We use a higher learning rate for the FC layers to enable such layers to learn more specific features in their respective domains, with respect to the rest of the model which is shared between source and target domains. We chose lr_{shared} by performing a grid search in the range $[2e^{-2}, 1e^{-4}]$, then we computed $lr_{FC} = 5 \cdot lr_{shared}$ [20]. We computed the best values on the ResNet18 and then used them also for the ViT. In both

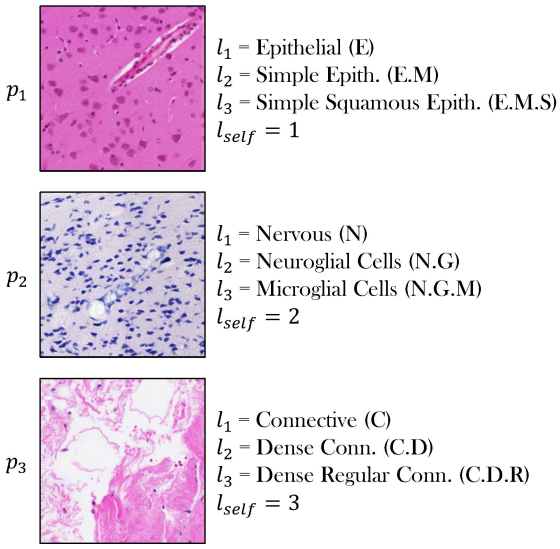


FIGURE 4. Examples of patches p belonging to the histopathology database Atlas of Digital Pathology (ADP). The associated set of labels are indicated on the right of each sample, with the label l_{i+1} representing a more precise classification than l_i. For simplicity, for each patch p and level l, only one type of tissue is indicated. Moreover, l_{self} indicates the progressive number of the WSI it was extracted from.

cases, we consider a deep tuning approach, enabling gradient update on all layers of the model [17]. To compute the losses we apply a class weighting procedure to compensate for the class imbalance, by weighing the contribution of each class to the loss by $w[i] = N/n_i$, where N is the total number of samples for the database and n_i is the number of samples for the $i-$th class (Eqn. 1, Eqn. 2). During the computation of the aggregated global loss, we chose the parameters w_1, w_2 by varying their values in the range $[0, 1]$ so that $w_1 + w_2 = 1$. We obtained the best results for $w_1 = w_2 = 0.5$. We also considered methods for automatically learning the best values [50], [51], without increasing the resulting accuracy. After the last epoch, we select the values of the weights for which we obtained the highest classification accuracy on the validation subset.

We train a different DL4ALL considering the two models (ResNet18 and ViT) and the three different learning procedures described in Section III-B, namely "regular", "greedy", and "self-supervised". For the "regular" and the "greedy" procedures, we consider the different labeling levels of the ADP database $L(p) = \{l_1, l_2, l_3\}$, as described in Section IV-A. Table 3 summarizes the different DL4ALL models obtained using the proposed approach.

C. ERROR MEASURES

We consider both quantitative and qualitative error measures. As quantitative error measures, we consider the classification accuracy describing the percentage of samples correctly classified, with respect to the total number of samples in the testing subset, the specificity, and the sensitivity. In addition, we include the confusion matrix, which describes the percentages of true positives, true negatives, false positives, and false negatives, according to the error measures described in [45].

As qualitative measures, we consider the t-Distributed Stochastic Neighbor Embedding (t-SNE) [52], since it is often used to show the distribution of the samples in the latent space, and the Grad-CAM [53], which outputs an activation map showing the degree in which the different regions of the image influence the model decision.

TABLE 3. Summary of the different DL4ALL models obtained using the proposed approach, based on the different learning procedures and the labels considered.

Procedure (Labels)	Model: ResNet18	Model: ViT
Regular (l_1)	DL4ALL $_{\text{reg,ResNet18,1}}$	DL4ALL $_{\text{reg,ViT,1}}$
Regular (l_2)	DL4ALL $_{\text{reg,ResNet18,2}}$	DL4ALL $_{\text{reg,ViT,2}}$
Regular (l_3)	DL4ALL $_{\text{reg,ResNet18,3}}$	DL4ALL $_{\text{reg,ViT,3}}$
Greedy (l_1)	DL4ALL $_{\text{greedy,ResNet18,1}}$	DL4ALL $_{\text{greedy,ViT,1}}$
Greedy (l_2)	DL4ALL $_{\text{greedy,ResNet18,2}}$	DL4ALL $_{\text{greedy,ViT,2}}$
Greedy (l_3)	DL4ALL $_{\text{greedy,ResNet18,3}}$	DL4ALL $_{\text{greedy,ViT,3}}$
Self-supervised	DL4ALL $_{\text{self,ResNet18}}$	DL4ALL $_{\text{self,ViT}}$

D. RESULTS

1) QUANTITATIVE EVALUATION

Table 4 shows the accuracy results, in terms of classification accuracy, specificity, and sensitivity, obtained using the proposed DL4ALL trained using the innovative cross-dataset transfer learning procedures, and applied on the CNMC database [13]. As a comparison, we include the corresponding results obtained using the ResNet18 and ViT in a standard transfer learning procedure with a deep tuning approach, in which the databases for the source and target domain are used in sequence, following a procedure often performed in medical imaging [2], [18]. We chose such procedures as a comparison since it allows us to compare the same backbone models under the different learning procedures. For example, ResNet18$_{ImageNet,ADP,CNMC}$ describes a ResNet18 pre-trained on the ImageNet database, then deep tuned on the ADP database, then again deep tuned and tested on the CNMC database. Moreover, we compare the results with the ALLNet and with the OrthoALLNet, methods based on the ResNet18 CNN that currently achieve the state of the art accuracy on the ALL-IDB2 and ALL-IDB Patches databases, respectively [19], [29]. Lastly, we report the results of the MMA-MTL model described in [15].

From Table 4, it is possible to observe that the DL4ALL trained using the proposed cross-dataset transfer learning procedures always increases the classification accuracy on the CNMC database. When comparing the proposed DL4ALL against the compared methods in a standard transfer learning approach, DL4ALL always performs equal or better, assuming the same model and the same level of labeling l_i on the source histopathological databases. As an example, DL4ALL based on the ResNet18 and using the l_1 labels of ADP performs better than the ResNet18 trained

TABLE 4. Accuracy results of DL4ALL on the CNMC database using the proposed methodology.

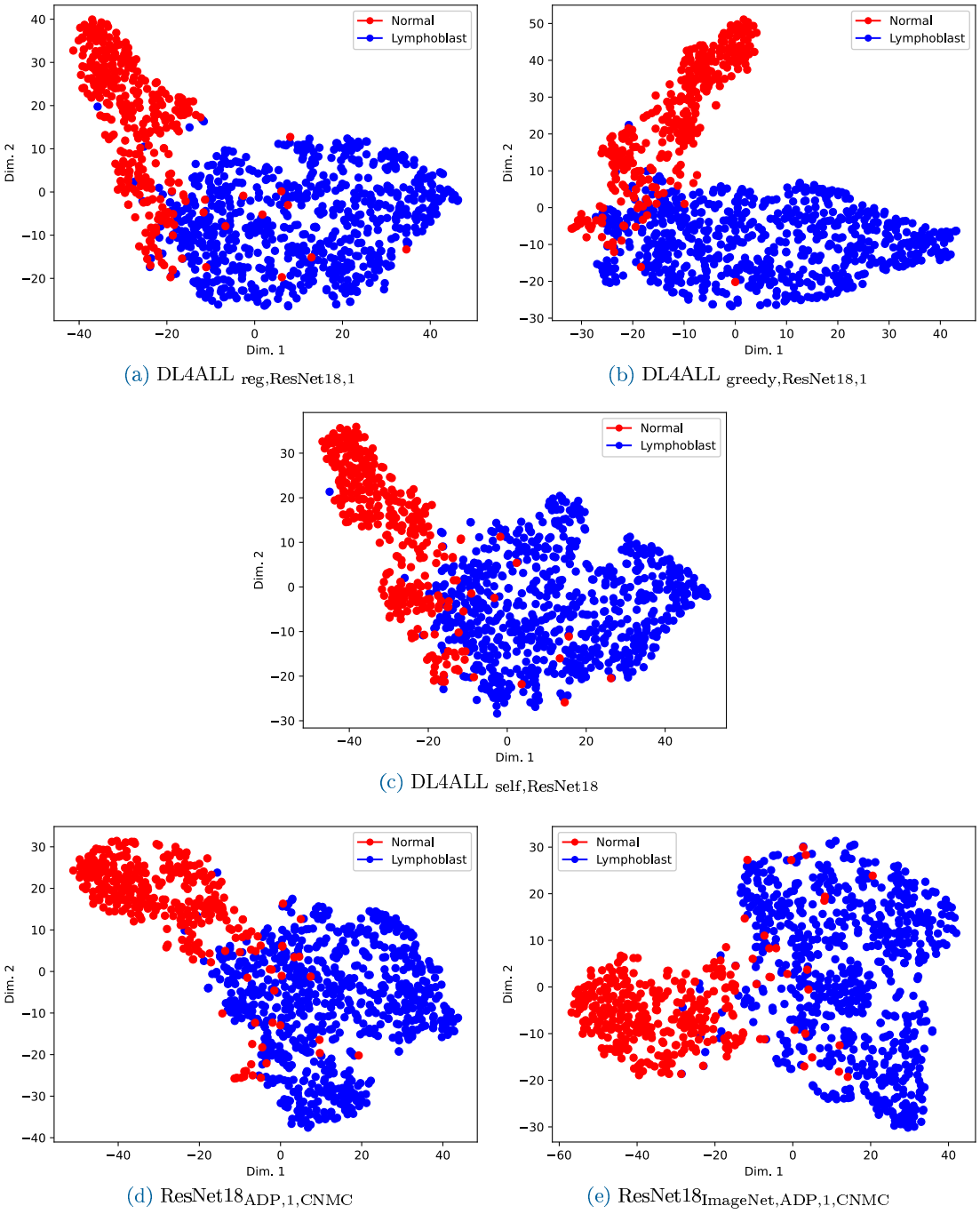
Ref.	Model	Class. Accuracy (%) (Mean_{Std})	Specificity	Sensitivity
[15]	MMA-MTL*	$93.85_{\text{N/A}}$	–	–
[2]	$\text{ResNet18}_{\text{ADP,1,CNMC}}$	$96.24_{0.81}$	$97.39_{0.80}$	$93.76_{2.35}$
	$\text{ResNet18}_{\text{ADP,2,CNMC}}$	$94.85_{0.36}$	$96.76_{1.03}$	$90.76_{1.39}$
	$\text{ResNet18}_{\text{ADP,3,CNMC}}$	$95.99_{0.16}$	$98.35_{0.54}$	$90.94_{1.33}$
	$\text{ResNet18}_{\text{ImageNet,ADP,1,CNMC}}$	$95.97_{0.36}$	$97.88_{0.41}$	$91.88_{1.70}$
	$\text{ResNet18}_{\text{ImageNet,ADP,2,CNMC}}$	$95.32_{1.08}$	$96.46_{0.90}$	$92.88_{1.78}$
	$\text{ResNet18}_{\text{ImageNet,ADP,3,CNMC}}$	$96.09_{0.76}$	$97.88_{0.62}$	$92.24_{2.72}$
[19]	$\text{ALLNet}_{\text{ADP,1,CNMC}}$	$92.98_{0.98}$	$96.84_{0.45}$	$84.71_{2.16}$
	$\text{ALLNet}_{\text{ADP,2,CNMC}}$	$93.54_{0.68}$	$96.07_{1.14}$	$88.12_{1.19}$
	$\text{ALLNet}_{\text{ADP,3,CNMC}}$	$92.53_{0.96}$	$96.07_{1.01}$	$84.94_{1.62}$
[29]	$\text{OrthoALLNet}_{\text{ADP,1,CNMC}}$	$94.03_{0.39}$	$97.80_{1.32}$	$85.94_{1.74}$
	$\text{OrthoALLNet}_{\text{ADP,2,CNMC}}$	$94.87_{0.58}$	$97.36_{0.36}$	$89.53_{1.40}$
	$\text{OrthoALLNet}_{\text{ADP,3,CNMC}}$	$95.15_{0.54}$	$97.03_{0.79}$	$91.12_{2.36}$
	$\text{OrthoALLNet}_{\text{ImageNet,ADP,1,CNMC}}$	$95.28_{0.47}$	$97.01_{1.01}$	$91.59_{1.84}$
	$\text{OrthoALLNet}_{\text{ImageNet,ADP,2,CNMC}}$	$95.07_{0.81}$	$96.81_{1.39}$	$91.35_{2.00}$
	$\text{OrthoALLNet}_{\text{ImageNet,ADP,3,CNMC}}$	$95.09_{0.46}$	$97.69_{0.32}$	$89.53_{1.64}$
-	$\text{ViT}_{\text{ADP,1,CNMC}}$	$89.29_{0.34}$	$95.58_{1.07}$	$75.82_{2.89}$
	$\text{ViT}_{\text{ADP,2,CNMC}}$	$90.02_{0.96}$	$95.11_{1.64}$	$79.12_{2.18}$
	$\text{ViT}_{\text{ADP,3,CNMC}}$	$80.11_{0.73}$	$91.54_{1.61}$	$55.65_{3.89}$
	$\text{ViT}_{\text{ImageNet,ADP,1,CNMC}}$	$97.10_{0.38}$	$98.54_{0.25}$	$94.00_{1.09}$
	$\text{ViT}_{\text{ImageNet,ADP,2,CNMC}}$	$96.67_{0.71}$	$98.41_{0.69}$	$92.94_{1.22}$
	$\text{ViT}_{\text{ImageNet,ADP,3,CNMC}}$	$91.65_{0.34}$	$95.47_{1.57}$	$83.47_{3.35}$
-	DL4ALL $_{\text{reg,ResNet18,1}}$	$96.24_{0.26}$	$97.18_{0.54}$	$94.20_{1.30}$
	DL4ALL $_{\text{reg,ResNet18,2}}$	$96.72_{0.49}$	$98.11_{0.51}$	$93.75_{2.33}$
	DL4ALL $_{\text{reg,ResNet18,3}}$	$96.61_{0.53}$	$98.16_{0.69}$	$93.27_{1.13}$
	DL4ALL $_{\text{greedy,ResNet18,1}}$	$97.02_{0.55}$	$97.64_{0.71}$	$95.69_{1.08}$
	DL4ALL $_{\text{greedy,ResNet18,2}}$	$96.95_{0.36}$	$97.94_{0.69}$	$94.80_{0.85}$
	DL4ALL $_{\text{greedy,ResNet18,3}}$	$96.91_{0.45}$	$97.86_{0.55}$	$94.87_{1.18}$
	DL4ALL $_{\text{self,ResNet18}}$	$96.20_{0.69}$	$97.21_{1.12}$	$94.03_{1.40}$
-	DL4ALL $_{\text{reg,ViT,1}}$	$97.38_{0.53}$	$98.33_{0.58}$	$95.33_{1.40}$
	DL4ALL $_{\text{reg,ViT,2}}$	$96.87_{0.34}$	$98.00_{0.72}$	$94.44_{1.31}$
	DL4ALL $_{\text{reg,ViT,3}}$	$96.85_{0.59}$	$98.41_{0.51}$	$93.51_{1.02}$
	DL4ALL $_{\text{greedy,ViT,1}}$	$97.25_{0.68}$	$98.77_{0.46}$	$94.00_{2.10}$
	DL4ALL $_{\text{greedy,ViT,2}}$	$\mathbf{97.85_{0.19}}$	$\mathbf{98.79_{0.18}}$	$\mathbf{95.81_{0.57}}$
	DL4ALL $_{\text{greedy,ViT,3}}$	$97.15_{0.69}$	$98.63_{0.38}$	$93.98_{2.39}$
	DL4ALL $_{\text{self,ViT}}$	$96.78_{0.54}$	$98.46_{0.38}$	$93.15_{1.52}$

Notes. * The work does not report the classification accuracy, so we reported the best error measure mentioned in the paper.

using the l_1 labels of ADP and fine tuned on CNMC, even when pretraining on ImageNet: DL4ALL $_{ResNet18,1} \geq \text{ResNet18}_{ADP,1,CNMC} \geq \text{ResNet18}_{ImageNet,ADP,1,CNMC}$. The increase in accuracy is valid also when comparing against ALLNet and OrthoALLNet, which are based on the ResNet18, and when considering both the "regular" and the "greedy" learning algorithms, indicating that the proposed methodology can effectively increase the classification accuracy on the target database, with respect to using the standard transfer learning approach currently used for ALL detection, by considering the source and target domains in an interleaved fashion with a multi-task architecture. In particular, DL4ALL $_{greedy,ViT,2}$ achieves the highest accuracy, highest specificity, and highest sensitivity on the CNMC database, with in average 97.85 % of correctly classified samples, a specificity of 98.79%, and a sensitivity of 95.81%, indicating the validity of the proposed approach. Moreover, Table 5 reports the corresponding confusion matrix.

It is worth noting that the "self-supervised" algorithm of DL4ALL, both using the ResNet18 and the ViT architectures, performs almost in-par with the state of the art, assuming the same architecture, despite not using manually-obtained labels at the sample level for the source domain. This result indicate a limited necessity for labeled samples in the source domain, possibly enabling the collection of even larger histopathological databases, when removing the necessity for a complex and time-consuming labeling of each patch, in turn fostering the research on further transfer learning approaches leveraging larger databases in the source domain.

We performed the experiments on a machine composed of an Intel Core i7 7800X @4 GHz CPU, 32 GB DDR4 RAM @ 2667 MHz, SSD SATA PCIe 256 GB, 2 x NVIDIA



(a) DL4ALL $_{\text{reg,ResNet18,1}}$ (b) DL4ALL $_{\text{greedy,ResNet18,1}}$

(c) DL4ALL $_{\text{self,ResNet18}}$

(d) ResNet18$_{\text{ADP,1,CNMC}}$ (e) ResNet18$_{\text{ImageNet,ADP,1,CNMC}}$

FIGURE 5. Result of the t-SNE algorithm on the feature space corresponding to the last convolutional layer of DL4ALL and analyzing the two classes of the CNMC database (*a*–*c*), compared with a ResNet18 trained using a standard transfer learning procedure (*d*–*e*). It is possible to see how the proposed approach better disentangles the representations of the feature space, when classifying CNMC samples as either "normal" or "lymphoblast".

RTX 6000, running PyTorch. The ViT model in the ''greedy'' configuration (DL4ALL $_{greedy,ViT}$) took the longest to train, with an average of 36 hours of training for 100 epochs. This is due to the high number of parameters of the ViT ($\approx 86M$) and the fact that the ''greedy'' learning algorithm has to perform additional computations to roll back the weight update step for every time it processes a batch for which the accuracy does not increase (see Fig. 3). Because of the considerable training times, we did not perform a tuning of the hyperparameters (e.g., the learning rate) for the ViT, which may result in even higher accuracies. Instead, we computed the best values on the ResNet18 and then used them also for the ViT.

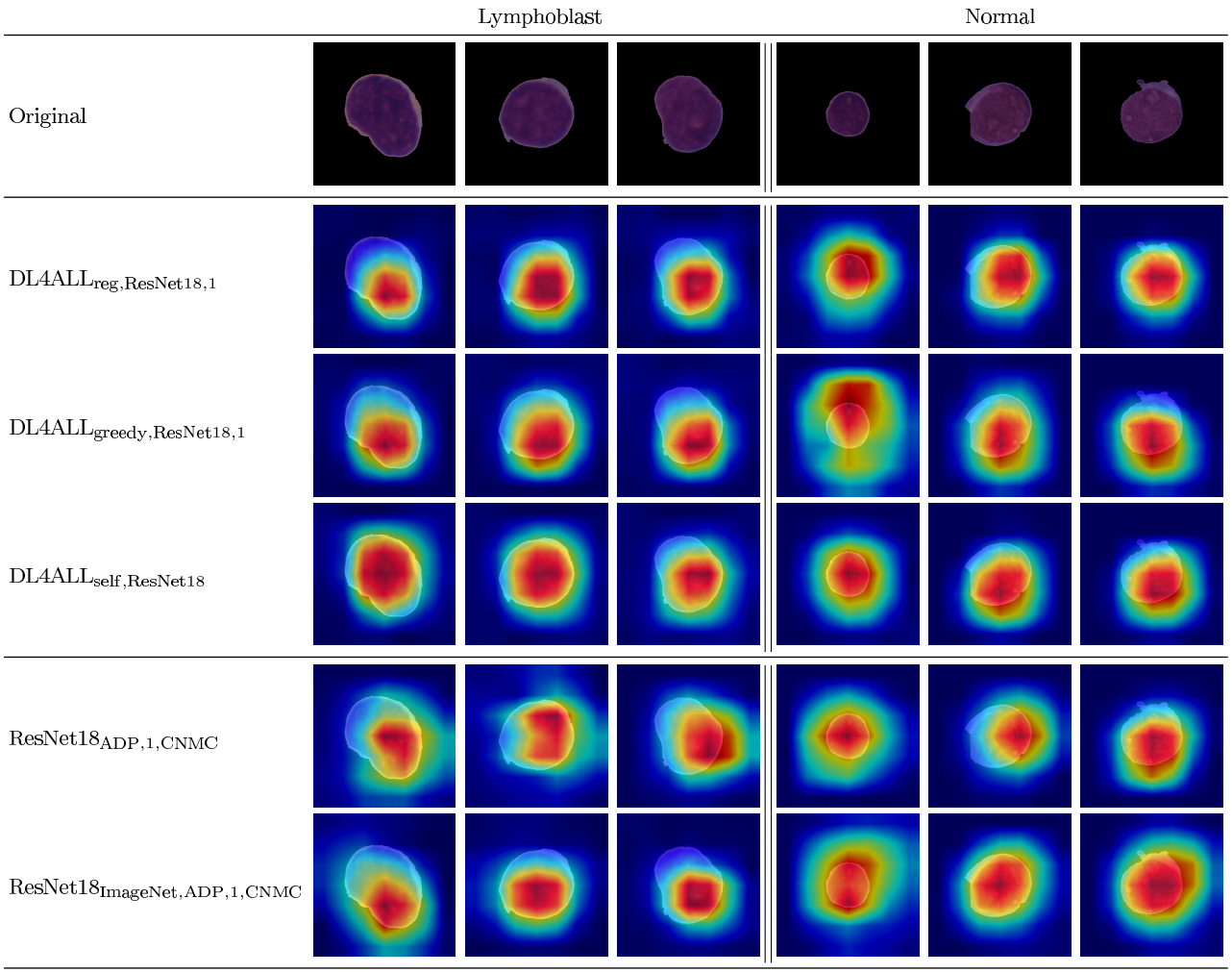


FIGURE 6. Result of the Grad-CAM method considering the DL4ALL models trained using the proposed methodology and applied on the CNMC database. It is possible to see that the heatmaps obtained using the $DL4ALL_{self,ResNet18}$ are the most focused on the WBC themselves, rather than the background.

However, in this paper we considered an unofficial implementation of the ViT,[4] not included in the PyTorch distribution. Future optimized implementations may reduce this gap.

2) QUALITATIVE EVALUATION

To perform the qualitative evaluation we considered our approaches based on CNNs, since the t-SNE and Grad-CAM algorithms are mostly used for analyzing the predictions of CNN-based models.

We apply the t-SNE algorithm [52] on the feature space corresponding to the last convolutional layer of $DL4ALL_{reg,ResNet18,1}$, $DL4ALL_{greedy,ResNet18,1}$, and $DL4ALL_{self,ResNet18}$ and considering the two classes in the CNMC database. We considered the methods using *ResNet*18, 1 since the $DL4ALL_{greedy,ResNet18,1}$ is the best performing of the CNN-based approaches. We compare the results with the state of the art represented by $ResNet18_{ADP,1,CNMC}$ and $ResNet18_{ImageNet,ADP,1,CNMC}$.

[4] https://github.com/lukemelas/PyTorch-Pretrained-ViT

TABLE 5. Average confusion matrix of the $DL4ALL_{greedy,ViT,2}$ on the CNMC database using the proposed methodology.

		Predicted	
		0 (*normal*)	1 (*lymphoblast*)
True	0 (*normal*)	TN = 67.60%	FP = 0.74%
	1 (*lymphoblast*)	FN = 1.12%	TP = 30.52%

Notes. TN = True Negatives; TP = True Positives; FN = False Negatives; FP = False Positives.

Fig. 5 shows the results. From the Figure, it is possible to observe that the proposed approach better disentangles the representations of the feature space, when classifying CNMC

samples as either ''normal'' or ''lymphoblast'', with respect to the state of the art.

We also apply the Grad-CAM method, as shown in Fig. 6. From the Figure, it is possible to observe that the heatmaps obtained using the DL4ALL trained with the proposed methodology are more focused on the WBCs, suggesting that DL4ALL is able to learn features more related to the cell itself, rather than database-specific (e.g., the background). In particular, the Grad-CAM heatmaps obtained using DL4ALL $_{self,ResNet18}$ are the ones most corresponding to each WBC.

V. CONCLUSION

In this paper we proposed a method named ''Deep Learning for Acute Lymphoblastic Leukemia'' (DL4ALL), the first approach in the literature based on multi-task learning and cross-dataset transfer learning for Acute Lymphoblastic (or Lymphocytic) Leukemia (ALL) detection. The method first adapts an existing deep model into a multi-task learning configuration, then uses innovative learning procedures that consider databases from both the source and target domains at the same time, even when they are significantly different, by interleaving batches during training. We proposed different variations of the cross-dataset transfer learning procedure, namely ''regular'', ''greedy'', and ''self-supervised'', based on how the source domain and the corresponding labels are used to help increase the classification accuracy of the target domain.

The results on a publicly-available ALL dataset demonstrate a greater accuracy in detecting ALL samples with respect to current methods in the literature. In particular, the ''greedy'' learning procedure, which at each iteration considers the source domain only if it helps in reducing the target loss, achieved the best results. Moreover, the ''self-supervised'' also performed better than the state of the art, despite not using manually-obtained labels at the sample level for the source domain. Future works will consider different databases as either source or target domains, different DL models, different learning algorithms, as well as more computationally efficient implementations, and optimized architectures.

ACKNOWLEDGMENT

The authors thank NVIDIA Corporation for the GPU donated. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the Italian MUR. Neither the European Union nor Italian MUR can be held responsible for them.

REFERENCES

[1] A. Genovese, M. S. Hosseini, V. Piuri, K. N. Plataniotis, and F. Scotti, ''Acute lymphoblastic leukemia detection based on adaptive unsharpening and deep learning,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 1205–1209.

[2] A. Genovese, M. S. Hosseini, V. Piuri, K. N. Plataniotis, and F. Scotti, ''Histopathological transfer learning for acute lymphoblastic leukemia detection,'' in *Proc. IEEE Int. Conf. Comput. Intell. Virtual Environments Meas. Syst. Appl. (CIVEMSA)*, Jun. 2021, pp. 1–6.

[3] S. Kermani, M. Amin, and A. Talebi, ''Recognition of acute lymphoblastic leukemia cells in microscopic images using K-means clustering and support vector machine classifier,'' *J. Med. Signals Sensors*, vol. 5, no. 1, p. 49, 2015.

[4] A. Mittal, S. Dhalla, S. Gupta, and A. Gupta, ''Automated analysis of blood smear images for leukemia detection: A comprehensive review,'' *ACM Comput. Surveys*, vol. 54, no. 11, pp. 1–37, Jan. 2022.

[5] M. Zolfaghari and H. Sajedi, ''A survey on automated detection and classification of acute leukemia and WBCs in microscopic blood cells,'' *Multimedia Tools Appl.*, vol. 81, no. 5, pp. 6723–6753, Feb. 2022.

[6] A. S. Panayides, A. Amini, N. D. Filipovic, A. Sharma, S. A. Tsaftaris, A. Young, D. Foran, N. Do, S. Golemati, T. Kurc, K. Huang, K. S. Nikita, B. P. Veasey, M. Zervakis, J. H. Saltz, and C. S. Pattichis, ''AI in medical imaging informatics: Current challenges and future directions,'' *IEEE J. Biomed. Health Informat.*, vol. 24, no. 7, pp. 1837–1857, Jul. 2020.

[7] J. Du, K. Guan, Y. Zhou, Y. Li, and T. Wang, ''Parameter-free similarity-aware attention module for medical image classification and segmentation,'' *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 3, pp. 1–13, Jun. 2022.

[8] M. A. Alsalem, A. A. Zaidan, B. B. Zaidan, M. Hashim, H. T. Madhloom, N. D. Azeez, and S. Alsyisuf, ''A review of the automated detection and classification of acute leukaemia: Coherent taxonomy, datasets, validation and performance measurements, motivation, open challenges and recommendations,'' *Comput. Methods Programs Biomed.*, vol. 158, pp. 93–112, May 2018.

[9] H. T. Salah, I. N. Muhsen, M. E. Salama, T. Owaidah, and S. K. Hashmi, ''Machine learning applications in the diagnosis of leukemia: Current trends and future directions,'' *Int. J. Lab. Hematology*, vol. 41, no. 6, pp. 717–725, Dec. 2019.

[10] S. Shafique and S. Tehsin, ''Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks,'' *Technol. Cancer Res. Treatment*, vol. 17, Jan. 2018, Art. no. 153303381880278.

[11] A. Rehman, N. Abbas, T. Saba, S. I. U. Rahman, Z. Mehmood, and H. Kolivand, ''Classification of acute lymphoblastic leukemia using deep learning,'' *Microsc. Res. Technique*, vol. 81, no. 11, pp. 1310–1317, Nov. 2018.

[12] A. T. Sahlol, P. Kollmannsberger, and A. A. Ewees, ''Efficient classification of white blood cell leukemia with improved swarm optimization of deep features,'' *Sci. Rep.*, vol. 10, no. 1, p. 2536, Feb. 2020.

[13] R. Duggal, A. Gupta, R. Gupta, and P. Mallick, ''SD-Layer: Stain deconvolutional layer for CNNs in medical microscopic imaging,'' in *Proc. MICCAI*, 2017, pp. 435–443.

[14] J. L. Wang, A. Y. Li, M. Huang, A. K. Ibrahim, H. Zhuang, and A. M. Ali, ''Classification of white blood cells with PatternNet-fused ensemble of convolutional neural networks (PECNN),'' in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2018, pp. 325–330.

[15] P. Mathur, M. Piplani, R. Sawhney, A. Jindal, and R. R. Shah, ''Mixup multi-attention multi-tasking model for early-stage leukemia identification,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1045–1049.

[16] B. Masoudi, ''VKCS: A pre-trained deep network with attention mechanism to diagnose acute lymphoblastic leukemia,'' *Multimedia Tools Appl.*, vol. 82, no. 12, pp. 18967–18983, Nov. 2022.

[17] R. Zhang, J. Zhu, S. Yang, M. S. Hosseini, A. Genovese, L. Chen, C. Rowsell, S. Damaskinos, S. Varma, and K. N. Plataniotis, ''Histokt: Cross knowledge transfer in computational pathology,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 1276–1280.

[18] H. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, ''Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning,'' *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[19] A. Genovese, "ALLNet: Acute lymphoblastic leukemia detection using lightweight convolutional networks," in *Proc. IEEE 9th Int. Conf. Comput. Intell. Virtual Environments Meas. Syst. Appl. (CIVEMSA)*, Jun. 2022, pp. 1–6.

[20] R. Mormont, P. Geurts, and R. Marée, "Multi-task pre-training of deep neural networks for digital pathology," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 2, pp. 412–421, Feb. 2021.

[21] X. Yao, Z. Zhu, C. Kang, S. Wang, J. M. Gorriz, and Y. Zhang, "AdaD-FNN for chest CT-based COVID-19 diagnosis," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 1, pp. 5–14, Feb. 2023.

[22] M. Chen, K. Q. Weinberger, and J. Blitzer, "Co-training for domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds. vol. 24, 2011, pp. 1–9.

[23] T. Isobe, X. Jia, S. Chen, J. He, Y. Shi, J. Liu, H. Lu, and S. Wang, "Multi-target domain adaptation with collaborative consistency learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8183–8192.

[24] Q. Liao, Y. Ding, Z. L. Jiang, X. Wang, C. Zhang, and Q. Zhang, "Multi-task deep convolutional neural network for cancer diagnosis," *Neurocomputing*, vol. 348, pp. 66–73, Jul. 2019.

[25] Y. Ganin, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, May 2016.

[26] M. S. Hosseini, L. Chan, W. Huang, Y. Wang, D. Hasan, C. Rowsell, S. Damaskinos, and K. N. Plataniotis, "On transferability of histological tissue labels in computational pathology," in *Proc. ECCV*, 2020, pp. 453–469.

[27] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[28] A. Loddo and L. Putzu, "On the effectiveness of leukocytes classification methods in a real application scenario," *AI*, vol. 2, no. 3, pp. 394–412, Aug. 2021.

[29] A. Genovese, V. Piuri, and F. Scotti, "ALL-IDB patches: Whole slide imaging for acute lymphoblastic leukemia detection using deep learning," in *Proc. ICASSP*, 2023, pp. 1–5.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[31] M. E. Billah and F. Javed, "Bayesian convolutional neural network-based models for diagnosis of blood cancer," *Appl. Artif. Intell.*, vol. 36, no. 1, pp. 1–22, Dec. 2022.

[32] A. Kumar, J. Rawat, I. Kumar, M. Rashid, K. U. Singh, Y. D. Al-Otaibi, and U. Tariq, "Computer-aided deep learning model for identification of lymphoblast cell using microscopic leukocyte images," *Exp. Syst.*, vol. 39, no. 4, May 2022, Art. no. e12894.

[33] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M.-L. Shyu, S.-C. Chen, and S. S. Iyengar, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–36, Sep. 2018.

[34] M. Abukmeil, S. Ferrari, A. Genovese, V. Piuri, and F. Scotti, "A survey of unsupervised generative models for exploratory data analysis and representation learning," *ACM Comput. Surveys*, vol. 54, no. 5, pp. 1–40, Jun. 2022.

[35] (2023). *Pytorch*. [Online]. Available: https://pytorch.org/

[36] J. Zhao, X. Qian, Y. Zhang, D. Shan, X. Liu, S. Coleman, and D. Kerr, "A knowledge distillation-based multi-scale relation-prototypical network for cross-domain few-shot defect classification," *J. Intell. Manuf.*, pp. 1–17, Feb. 2023. [Online]. Available: https://link.springer.com/article/10.1007/s10845-023-02080-w

[37] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.

[38] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surveys*, vol. 53, no. 3, pp. 1–34, Jun. 2020.

[39] J. Long, Y. Chen, Z. Yang, Y. Huang, and C. Li, "A novel self-training semi-supervised deep learning approach for machinery fault diagnosis," *Int. J. Prod. Res.*, pp. 1–14, Feb. 2022. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/00207543.2022.2032860

[40] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," in *Neural Networks: Tricks of the Trade*, 2nd ed. Berlin, Germany: Springer, 2012, pp. 599–619.

[41] X. Zheng, Y. Wang, G. Wang, and J. Liu, "Fast and robust segmentation of white blood cell images by self-supervised learning," *Micron*, vol. 107, pp. 55–71, Apr. 2018.

[42] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[43] T. Ridnik, E. B. Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21K pretraining for the masses," 2021, *arXiv:2104.10972*.

[44] M. S. Hosseini, L. Chan, G. Tse, M. Tang, J. Deng, S. Norouzi, C. Rowsell, K. N. Plataniotis, and S. Damaskinos, "Atlas of digital pathology: A generalized hierarchical histological tissue type-annotated database for deep learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11739–11748.

[45] R. D. Labati, V. Piuri, and F. Scotti, "All-IDB: The acute lymphoblastic leukemia image database for image processing," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 2045–2048.

[46] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd ed. Cambridge, MA, USA: MIT Press, 2001.

[47] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021.

[48] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[49] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1–7.

[50] L. Liebel and M. Körner, "Auxiliary tasks in multi-task learning," 2018, *arXiv:1805.06334*.

[51] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," in *Proc. NIPS*, 2018, pp. 525–536.

[52] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[53] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

ANGELO GENOVESE (Senior Member, IEEE) received the Ph.D. degree in computer science from the Università degli Studi di Milano, Italy, in 2014.

Since 2022, he has been an Associate Professor in computer science with the Università degli Studi di Milano. He has been a Visiting Researcher at the University of Toronto, Toronto, ON, Canada. His original results have been published in more than 70 papers in international journals, proceedings of international conferences, books, and book chapters. His research interests include signal and image processing, 3-D reconstruction, artificial intelligence for industrial and environmental monitoring systems, biometric systems, and design methodologies and algorithms for self-adapting systems.

Dr. Genovese is an Associate Editor of the *Journal of Ambient Intelligence and Humanized Computing* (Springer).

VINCENZO PIURI (Fellow, IEEE) received the Ph.D. degree in computer engineering from the Politecnico di Milano, Milan, Italy, in 1989.

He has been a Full Professor of computer engineering at the Università degli Studi di Milano, Milan, since 2000. He has also been an Associate Professor at the Politecnico di Milano and a Visiting Professor at The University of Texas at Austin, USA, and a Visiting Researcher at George Mason University, USA. He is an Honorary Professor at Obuda University, Hungary; Guangdong University of Petrochemical Technology, China; Northeastern University, China; Muroran Institute of Technology, Japan; and Amity University, India. His research interests include artificial intelligence, computational intelligence, intelligent systems, machine learning, pattern analysis and recognition, signal and image processing, biometrics, intelligent measurement systems, industrial applications, digital processing architectures, fault tolerance, dependability, and cloud computing infrastructures. His original results have been published in more than 400 articles in international journals, proceedings of international conferences, books, and book chapters.

Dr. Piuri is a Distinguished Scientist of ACM and a Senior Member of INNS. He is the President of the IEEE Systems Council (2020–2021) and has been the IEEE Vice President for Technical Activities (2015), an IEEE Director, the President of the IEEE Computational Intelligence Society, the Vice President for Education of the IEEE Biometrics Council, the Vice President for Publications of the IEEE Instrumentation and Measurement Society and the IEEE Systems Council, and the Vice President for Membership of the IEEE Computational Intelligence Society. He received the IEEE Instrumentation and Measurement Society Technical Award, in 2002, and the IEEE TAB Hall of Honor, in 2019. He has been the Editor-in-Chief of IEEE Systems Journal (2013–2019) and an Associate Editor of IEEE Transactions on Cloud Computing. He has been an Associate Editor of IEEE Transactions on Computers, IEEE Transactions on Neural Networks, IEEE Transactions on Instrumentation and Measurement, and IEEE Access.

KONSTANTINOS N. PLATANIOTIS (Fellow, IEEE) is currently a Professor and the Bell Canada Chair in multimedia with the ECE Department, University of Toronto. He is a Registered Professional Engineer in Ontario and a fellow of the Engineering Institute of Canada. He was the Technical Co-Chair for ICASSP 2013, the General Co-Chair for 2017 IEEE GlobalSIP, and Co-Chair for the 2018 IEEE International Conference on Image Processing (ICIP 2018). He is the General Co-Chair of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021). He has served as Signal Processing Society Vice President for Membership (2014–2016) and the Editor-in-Chief for IEEE Signal Processing Letters (2009–2011).

FABIO SCOTTI (Senior Member, IEEE) received the Ph.D. degree in computer engineering from the Politecnico di Milano, Milan, Italy, in 2003.

He has been a Full Professor at the Università degli Studi di Milano, Milan, since 2020. His original results have been published in over 130 papers in international journals, proceedings of international conferences, books, book chapters, and patents. His current research interests include biometric systems, machine learning and computational intelligence, signal and image processing, theory and applications of neural networks, 3-D reconstruction, industrial applications, intelligent measurement systems, and high-level system design.

Dr. Scotti is an Associate Editor of the IEEE Transactions on Systems, Man, and Cybernetics and the IEEE Open Journal of Signal Processing. He is serving as a Book Editor (an Area Editor, section less-constrained biometrics) for the *Encyclopedia of Cryptography, Security, and Privacy* (3rd Edition, Springer). He has been an Associate Editor of the IEEE Transactions on Information Forensics and Security and *Soft Computing* (Springer) and a Guest Co-Editor of the IEEE Transactions on Instrumentation and Measurement.

Open Access funding provided by 'Università degli Studi di Milano' within the CRUI CARE Agreement