A Knowledge-Reserved Distillation with Complementary Transfer for Automated FC-based Classification Across Hematological Malignancies

Jeng-Lin Li[1], Ting-Yu Chang[2], Yu-Fen Wang[2], Bor-Sheng Ko[2], Jih-Luh Tang[3] and Chi-Chun Lee[1]

Abstract— **Acute leukemia often comes with life-threatening prognosis outcome and remains a critical clinical issue today. The implementation of measurable residual disease (MRD) using flow cytometry (FC) is highly effective but the interpretation is time-consuming and suffers from physician idiosyncrasy. Recent machine learning algorithms have been proposed to automatically classify acute leukemia samples with and without MRD to address this clinical need. However, most prior works either validate only on a small data cohort or focus on one specific type of leukemia which lacks generalization. In this work, we propose a transfer learning approach in performing automatic MRD classification that takes advantage of a large scale acute myeloid leukemia (AML) database to facilitate better learning on a small cohort of acute lymphoblastic leukemia (ALL). Specifically, we develop a knowledge-reserved distilled AML pre-trained network with ALL complementary learning to enhance the ALL MRD classification. Our framework achieves 84.5% averaged AUC which shows its transferability across acute leukemia, and our further analysis reveals that younger and elder ALL patient samples benefit more from using the pre-trained AML model.**

I. INTRODUCTION

Acute leukemia is a fatal hematological malignancy that contains various subtypes, such as acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). ALL often occurs in childhood whereas AML is the most common leukemia in adults. While recent medical advancements have improved the survival rates for patients of leukemia, around 40%-50% of adult ALL patients still suffer from abrupt relapse [1]. This type of unexpected prognosis outcomes impose high risk on these patients. Clinically, identification of minimum residual disease (MRD) is proven to be a prognosis management indicator that helps stratify risk of therapy and guides treatment decisions [2]. Hence, detecting MRD is regularly administrated as a standard clinical examination using flow cytometry (FC) technology not only for diagnosis but also as a follow-up treatment efficacy indicator.

In the FC machine, each single cell is incubated with a panel of antibody markers through fluorescent excitement. The quantification signal corresponding to each antigen expression results in high dimensional data recorded in the FC output. The current practice relies on physicians to conduct manual gating by visualizing two-dimensional scatter plots of antibody-fluorescence pairs for MRD clinical interpretation. This complex and combinatorial pairing results in an extreme laborious and time-consuming interpretation of FC data. To improve the efficiency, several FC-based classification algorithms have been developed, e.g., Reiter et al. used a GMM approach to identify MRD in childhood acute Blymphoblastic leukemia in a 337-sample database [3]. Ko et al. proposed to use a specimen-level cytometric phenotype vector to differentiate AML and myelodysplastic syndrome (MDS) from the ones without MRD in a large scale database [4]. These past works [4], [5], while developed and validated on a large quantity of samples, they only concentrated on FC data with AML panel. Other automated ALL MRD identification works are only validated on a relatively small cohort [3], [6], which is likely due to the difficulty in collecting the same quantity of samples as AML. In this work, we aim at leveraging the AML MRD recognition model trained on large quantity to facilitate training of ALL MRD classification model, which has a smaller data available.

This scenario, i.e., learning from a large quantity of related samples to enhance the target model's capacity with smaller sample size, is often realized through the use of transfer learning. Recently, approaches of knowledge distillation, a.k.a., teacher student learning, has become the state-of-the-art transfer learning method in guiding student network learning by constraining the prediction to be similar to the teacher network [7]. However, a key shortcoming of this strategy is that the student network cannot surpass the teacher network's capability, which limits its applicability to transfer between related but not identical recognition tasks.

In our setup, AML and ALL while both are hematological malignancies, they remain two different disease types with its unique pathological characteristics. In order to transfer the model learned using AML samples to aid ALL classification model, teacher student training alone can be insufficient. In this work, we further employ complementary learning strategy to enhance the target (ALL) model capacity by explicitly capturing residual beyond source (AML) model to improve the cross-task training. A similar strategy has been evaluated in computer vision tasks [8]. Hence, in this work, we specifically propose a learning strategy of knowledge-reserved distillation with complementary transfer to facilitate the model training of ALL from a pre-trained AML model.

Our proposed framework demonstrates an improved classification accuracy compared to training directly only on ALL samples without transfer. Specifically, our model achieves an

[1] CCL, JLL are with Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan, and Joint Research Center for AI Technology and All Vista Healthcare, Ministry of Science and Technology, Taiwan (phone: +88635162439. e-mail: cclee@ee.nthu.edu.tw, cllee@gapp.nthu.edu.tw).

[2] TYC, YFW, BSK are with Department of Internal Medicine, College of Medicine, National Taiwan University, Taipei, Taiwan.

[3] BSK is with Division of Hematology, Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan.

[4] BSK, JLT are with Department of Hematological Oncology, National Taiwan University Cancer Center, Taipei, Taiwan.

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

TABLE I

DEMOGRAPHICS OF THE ANALYZED DATABASES

	Gender		Age			
Samples	Female	Male	0-15	15-39	39-65	65-
ALL	1273	1082	547	1002	726	74
AML	2239	2131	206	1383	2161	599

TABLE II

A LIST OF FLUORESCENT-MARKER IN AML AND ALL FC PANEL

Shared	FITC	CD5,Anti-HLA-DR,CD2,CD34,CD56
	PE	CD38,CD34, CD19,CD117,CD13,CD11b
	PerCP	CD45
AML	FITC	CD16,CD20,CD22,CD33
	PE	CD10
ALL	FITC	CD14,CD15,CD7
	PE	CD56,CD33

average 84.5% AUC, which outperforms other fine-tuning and knowledge distillation methods. Further analysis shows that our transfer learning strategy improves the most on younger and elder patient groups. An additional advantage of our proposed model-based transfer strategy is its ability to benefit ALL MRD classification in an AML data free condition, which further addresses the growing privacy issue of medical data sharing between sites.

II. METHODS

A. Database

The database for this study is collected from 2009 to 2016 from the National Taiwan University Hospital. These retrospective FC data samples comprise of patients who went through bone marrow aspiration. The specimen were originally examined by two flow cytometry machines (FASCalibur and FASCanto from Becton Dickinson Bioscience). Two panels of fluorescent markers were used for ALL and AML clinical diagnoses respectively. These specimens were investigated by certified physicians using the standard manual hierarchical gating procedure. The physicians labeled the data with MRD as "AML" or "ALL" depending on the targeted panel disease type. Thus, each panel of data was categorized into "positive MRD" or "no MRD (normal)" where "no MRD" indicated the specimen has no residual leukemia cells in terms of the examined panel disease. The data using ALL panel contains 493 patients which results in a total of 2356 unique bone marrow specimen samples (279 ALL, 720 normal in Calibur machine; 355 ALL, 1002 normal in Canto Machine). On the other hand, the data using AML panel has 1629 patients which includes 4372 samples in total (597 AML, 1564 normal in Calibur machine; 538 AML, 1673 normal in Canto Machine). This study is approved by the institutional ethics committee of National Taiwan University Hospital (No. 201906018RINB).

B. Cytometric Phenotype Embedding

We first encode each FC data sample to a fixed length cytometric phenotype embedding at the specimen level following our previous work [4]. Each FC data contains 100,000 cells, and we gather all the combinations of fluorescence-marker pairs for each cell to learn a vectorized representation at the specimen level based on a Gaussian Mixture Model (GMM) Fisher-scoring encoding approach. Given the cell level flow cytometry data $X = x_t$ where $t \in T$ (T indicates the total number of cells). We first learn a GMM probability function with parameters λ including weights, mean vectors, and covariance matrix, denoted as w_k, μ_k, and Σ_k for the k-th cluster of GMM. The final vector is derived using Fisher scoring function, which is the gradient of parameter λ. This results in a 2*K*D dimensions vector where D is the number of fluorescence-marker pairs and K is the number of GMM mixture. The specimen level representation is L2-normalized. We set mixture K as 16 through simple grid search.

C. Knowledge Distillation with Complementary Transfer

Our overall transfer learning framework is shown in Fig. 1. Our model-based transfer from AML MRD classification to ALL task is divided into two parts: distilling pre-trained AML model to derive knowledge-reserved network and complementary learning that adapts to ALL prediction task.

1) Knowledge-Reserved Network: We pre-train a deep neural network to classify AML MRD samples versus normal (no MRD) given the larger quantity of AML database available (shown in section II-A). Then, by training a separate classification network using ALL data, we specify KL divergence L_{KL} as a loss to keep the distribution of hidden layers in this ALL network to be similar to the pre-trained AML MRD network to perform distillation:

$$L_{KL} = \sum_{x \in X} p(x) * \log \frac{p(x)}{q(x)}$$

where p and q are the hidden layer outputs of the ALL and the AML network, respectively. We expect that the predictive behavior of this knowledge-reserved network to inherent discriminative power from the AML pre-trained model.

2) Complementary Transfer Learning: We further devise a complementary learning network that predicts the residual values, i.e., the difference between ALL ground truth and outputs of the knowledge-reserved network. Specifically, we firstly construct another ALL predictive network without distillation to be the same architecture as the knowledge-reserved network. Then, we feed the concatenated outputs of this ALL predictive network and knowledge-reserved network into a 2-layer network that predicts the residual values. The final prediction ($O = O_K + R$) is the sum from the output from knowledge-reserved network (O_K) and residual value (R). The loss used in the complementary learning network includes a cross entropy loss L_C of this predicted output O to the ALL ground truth. The ALL preditive network is trained with a ALL prediction loss L_T to ensure it being ALL discriminatively-favorable. The ALL predictive network and the complementary learning layers are jointly optimized with the joint loss $L = L_C + L_T$.

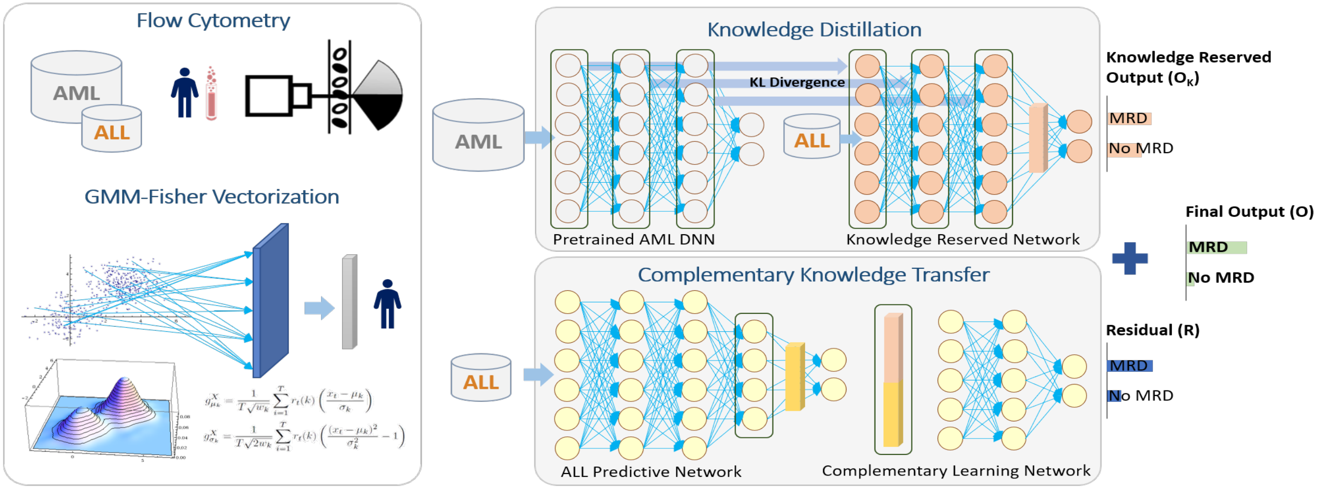

Fig. 1. The overall framework for knowledge-reserved distillation with complementary learning for ALL classification from pre-trained AML model.

D. Experimental Procedure

In this study, we examine the performance of classifying "positive MRD" versus "no MRD" in ALL FC panel using a 5-fold subject independent cross validation. Each fold includes 80% of specimens as training set and the rest 20% of specimens as testing set. Several evaluation metrics are used, including accuracy (ACC), unweighted average recall (UAR) and area under ROC curve (AUC). The optimized algorithm for all network training is based on Adam optimizer with relu activation function. The learning rate is 0.001 and batch size is 64. Three analyses are conducted in this study: comparison to other machine learning models, comparison to different transfer learning schemes, and age group analysis.

1) Comparison to Other Machine Learning Models: For each FC machine, we compare two typical machine learning (ML) methods to deep neural network (implemented using Pytorch toolbox):

- Logistic Regression (LR): using L2-regularization with strength 1.0.
- Support Vector Machine (SVM): using linear kernel with regularization parameter $C = 1$.
- Deep Neural Network (DNN): using a 3 fully-connected hidden layers followed by a softmax output layer. We specify 256 nodes for the layers and perform batch normalization before relu activation function. The final layer is a softmax layer that minimizes cross-entropy loss for binary classification labels.

2) Comparison to Different Transfer Learning Schemes:

- Pre-trained Network(PT): using AML pre-trained network directly without any fine-tuning on ALL data
- Fine-tune(FT): fine-tuning *DNN* that is initialized with the AML pre-trained model
- Knowledge distillation(KD): using knowledge-reserved network introduced in section II-C.1
- Fine-tune with complementary learning(FT-C): using *FT* model with complementary learning scheme described in section II-C.2.
- Knowledge-reserved distillation with complementary learning (KD-C): our proposed approach.

3) Age Group Analysis: A further analysis is carried out on the results of *KD-C* transfer learning model in terms of age groups. The analysis tries to identify which age groups benefit from such a AML to ALL transfer learning approach. This may help reveal the pathological relationship between ALL and AML characteristics as recorded in the FC data.

III. Results

In this study, we examine the "ALL" versus "no MRD" binary classification results using FC data collected from two FC machines (i.e., Canto and Calibur) which are shown in Table III. We observe that *KD-C*, i.e., our proposed approach, achieves consistently the highest performance as measured using ACC, UAR and AUC in both of the FC machines. When comparing between different baseline models without transfer, *SVM* attains a slightly higher performance than *LR* in Calibur machine (0.97% ACC, 4.18% AUC, 1.38% UAR relative high) whereas it performs moderately in Canto machine. *DNN* relatively improves 6.6% and 4.76 UAR compared to *SVM* in Calibur and Canto machine. Meanwhile, there are 2.81% and 1.98% relative AUC and ACC improvement in Canto. Generally, the *DNN* model shows its superior discriminative performance compared to other ML methods.

When examining accuracy obtained among transfer learning schemes, we observe that the baseline model *PT* already obtains accuracy beyond chance, which suggests that different subtypes of leukemia indeed share partial commonalities. The method of fine-tuning (*FT*) and knowledge distillation *KD* both achieve a better machine-averaged relative increase of ACC, AUC and UAR (i.e., 1.32%, 0.66%, 1.44% for *FT* and 1.02%, 1.98%, 2% for *KD*) when compared to *DNN*, i.e., training on ALL data only. This indicates the general issue that an inadequate amount of data would cause the performance to be sub-optimal. Fine-tuning from large scale AML pre-trained model provides a better initialization for ALL predictive network optimization which facilitates in finding a better convergent parameter space. *KD* can be thought as a soft regularization during network transfer learning to mitigate overfitting in the fine-tuning process on the small dataset.

TABLE III

Results of different ML algorithms and transfer learning approaches

	CantoII								Calibur							
	LR	SVM	DNN	PT	FT	KD	FT-C	KD-C	LR	SVM	DNN	PT	FT	KD	FT-C	KD-C
ACC	0.808	0.807	0.823	0.716	**0.841**	0.835	0.840	**0.841**	0.825	0.833	0.831	0.688	0.835	0.836	0.838	**0.847**
AUC	0.784	0.782	0.804	0.643	0.809	0.816	0.809	**0.819**	0.814	0.848	0.851	0.721	0.857	**0.872**	0.861	**0.870**
UAR	0.684	0.697	0.743	0.612	0.759	0.762	0.760	**0.771**	0.725	0.735	0.770	0.656	0.775	0.781	0.784	**0.787**

TABLE IV

Accuracy and relative improvement in different age groups

Age	0-15	15-39	39-65	65+	0-21	21-65	65+
DNN	0.806	0.847	0.809	0.851	0.811	0.832	0.851
KD-C	0.835	0.871	0.806	0.878	0.843	0.841	0.878
Δ (%)	**3.63**	2.83	-0.3	**3.17**	**4.03**	1.04	**3.17**

Furthermore, *FT-C* and *KD-C* are two models integrating complementary learning to further enhance the transfer efficacy. The machine-averaged relative improvements when compared to *DNN* come at 1.45% ACC, 0.87% AUC, and 2.1% UAR respectively when using *FT-C*, and 2.04% ACC, 2.08% AUC, 3.06% UAR when applying *KD-C*. We also observe that KD-C outperforms *FT-C*, demonstrating the importance in using the pre-trained AML network to perform knowledge-reserve distillation instead of simply use it as the initialization parameter as done in FT-based methods. The outstanding results of complementary learning is also attributed to the complex mechanism adopted in the additional residual prediction layers. Moreover, the knowledge distillation softly and continually regularizing the ALL network learning with its preserved AML knowledge is shown to be more appropriate when combining with complementary learning strategy; in contrast, fine-tuning AML model merely provides a better initialization point but has no involvement in the subsequent ALL model learning.

In our further analysis, we take the best-performing *KD-C* model and examine its improvement in ACC when compared to *DNN* across different age groups (the results are shown in Table IV). An obvious phenomenon is that the younger and the elder groups demonstrate a clear improvement when performing our proposed AML to ALL transfer. We report two group division results, i.e., [0,15,39,65] and [0,21,65]. The age group under 15 has 3.63% relative increase, age below 21 obtains an improvement of 4.03%, and the elders (age over 65) achieves 3.17% improvement.

IV. Discussion

Our experimental results show an encouraging transfer learning paradigm in classifying ALL from no MRD specimens by leveraging AML model as the source model. The proposed *KD-C* achieves the best 84.5% average AUC across two machines which outperforms *DNN* and other transfer learning schemes. An interesting observation to note is that when we investigate the age distribution in Table I in reference with the results shown in Table IV, while the young age patients is the majority in ALL data, the prediction on younger patients apparently still improved when using AML (where the adult is the majority) as the pre-trained model. Additionally, our proposed use of model-based transfer, i.e., no requirement in sharing the actual data samples, provides another advantage in securing the patient's data privacy when deploying this technology in real world across sites.

V. Conclusion

In this work, we propose a knowledge-reserved distillation with complementary learning strategy to facilitate ALL classification model training from a pre-trained AML model. We demonstrate advantages of our model-based transferability across acute leukemia subtypes implicating the potential toward leveraging large scale database of a specific subtype for other hematological malignancies without sharing the data itself. Further studies is necessary to examine the cross hematological malignancies relationship to help better understand the spectrum of these diseases and toward generalize the automated assistive solutions.

References

[1] A. K. Fielding, S. M. Richards, R. Chopra, H. M. Lazarus, M. R. Litzow, G. Buck, I. J. Durrant, S. M. Luger, D. I. Marks, I. M. Franklin, *et al.*, "Outcome of 609 adults after relapse of acute lymphoblastic leukemia (all); an mrc ukall12/ecog 2993 study," *Blood*, vol. 109, no. 3, pp. 944–950, 2006.

[2] I. Della Starza, S. Chiaretti, M. S. De Propris, L. Elia, M. Cavalli, L. A. De Novi, R. Soscia, M. Messina, A. Vitale, A. R. Guarini, *et al.*, "Minimal residual disease in acute lymphoblastic leukemia: technical and clinical advances," *Frontiers in oncology*, vol. 9, p. 726, 2019.

[3] M. Reiter, M. Diem, A. Schumich, M. Maurer-Granofszky, L. Karawajew, J. G. Rossi, R. Ratei, S. Groeneveld-Krentz, E. O. Sajaroff, S. Suhendra, *et al.*, "Automated flow cytometric mrd assessment in childhood acute b-lymphoblastic leukemia using supervised machine learning," *Cytometry Part A*, 2019.

[4] B.-S. Ko, Y.-F. Wang, J.-L. Li, C.-C. Li, P.-F. Weng, S.-C. Hsu, H.-A. Hou, H.-H. Huang, M. Yao, C.-T. Lin, *et al.*, "Clinically validated machine learning algorithm for detecting residual diseases with multicolor flow cytometry analysis in acute myeloid leukemia and myelodysplastic syndrome," *EBioMedicine*, vol. 37, pp. 91–100, 2018.

[5] J. Li, Y. Wang, B. Ko, C. Li, J. Tang, and C. Lee, "Learning a cytometric deep phenotype embedding for automatic hematological malignancies classification," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, July 2019, pp. 1733–1736.

[6] P. Rota, S. Groeneveld-Krentz, and M. Reiter, "On automated flow cytometric analysis for mrd estimation of acute lymphoblastic leukaemia: A comparison among different approaches," in *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov 2015, pp. 438–441.

[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[8] S. Wu, J. Li, C. Liu, Z. Yu, and H.-S. Wong, "Mutual learning of complementary networks via residual correction for improving semi-supervised classification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.