Learning a Cytometric Deep Phenotype Embedding for Automatic Hematological Malignancies Classification

Jeng-Lin Li[1], Yu-Fen Wang[2], Bor-Sheng Ko[3], Chi-Cheng Li[2,4], Jih-Luh Tang[2,3], Chi-Chun Lee[1]

***Abstract*— Identification of minimal residual disease (MRD) is important in assessing the prognosis of acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS). The current best clinical practice relies heavily on Flow Cytometry (FC) examination. However, the current FC diagnostic examination requires trained physicians to perform lengthy manual interpretation on high-dimensional FC data measurements of each specimen. The difficulty in handling idiosyncrasy between interpreters along with the time-consuming diagnostic process has become one of the major bottlenecks in advancing the treatment of hematological diseases. In this work, we develop an automatic MRD classifications (AML, MDS, normal) algorithm based on learning a deep phenotype representation from a large cohort of retrospective clinical data with over 2000 real patients' FC samples. We propose to learn a cytometric deep embedding through cell-level autoencoder combined with specimen-level latent Fisher-scoring vectorization. Our method achieves an average AUC of 0.943 across four different hematological malignancies classification tasks, and our analysis further reveals that with only half of the FC markers would be sufficient in obtaining these high recognition accuracies.**

I. INTRODUCTION

Acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS) are life-threatening hematologic malignancies with poor five year survival rate (less than 25%) [1]. Detection of minimal residual disease (MRD) that identifies leukemia cells from bone marrow is an effective indicator to trace follow-up condition of patients after completion of standard treatment [2]. In clinical practice, Flow Cytometry (FC), i.e., a leading technology enabling single cell monitoring, is one of the most prominent and clinically-validated tool to detect MRD for treatment outcome evaluation and therapeutic planning.

Each single cell of the blood sample drawn from the patient's bone marrow is first marked by a panel set of antibodies (a.k.a market set). Each cell's corresponding antigen expressions are further characterized by fluorescence values measured in different channels of the FC machine. Detection of MRD is then carried out through trained physician's lengthy manual interpretation of these *high-dimensional* FC data. Due to the high-dimensional measurements of a FC examination, each specimen has to undergo a complex and hierarchical manual gating procedure by visualizing multiple *two-dimensional* scatter plots, where each plot is generated from a set of antibody-FC channel combination. Not only is this manual interpretation procedure time consuming but also the idiosyncrasy existing among interpreters creates inevitable issues of reproducibility and objectivity. This current clinical practice hinders the efficiency of hematological malignancies diagnosis and treatment.

Recently, developing automated and reliable diagnosis and assessment techniques based on data-driven machine learning (ML) approaches have made significant progresses in the medical domain, e.g., stroke risk assessment [3], breast cancer diagnosis [4] and diabetic retinopathy detection [5]. Many of these ML approaches have relied on learning representation from the multi-variate clinical data measurements in order to achieve high recognition power. In the domain of hematology, few works have attempted to develop ML framework in detecting MRD from the FC measurements. Most of these works focus either on identification at the single cell-level (instead of specimen-level where the clinical diagnosis is made) or by learning from only limited real-world clinical samples. For example, Paolo et al. proposed a pattern recognition approach in identifying cancerous types at the cell-level [6]; Biehl et al. computed a variety of statistical functions to represent a specimen to perform ML-based classification [7], and Rajwa et al. developed an effective Bayesian-GMM based model to discriminate AML samples from normal ones although their database cohort included only 100 samples each [8].

In this work, we propose to learn a cytometric deep phenotype embedding vector to represent FC data sample at the specimen-level to perform automatic hematological malignancies classification (AML, MDS, and normal). Our database includes a large real patient cohort of 2424 FC data from the National Taiwan University Hospital (NTUH) gathered over the past five years. Our cytometric phenotype embedding is learned using an cell-level autoencoder with specimen-level latent-vector Fisher-scoring approach. It achieves a remarkable accuracy of 94.9%, 95.6%, 95.6%, and 91.1% in classifying FC samples for tasks of abnormal (AML+MDS) vs. normal, AML vs. normal, MDS vs. normal, and AML vs. MDS, respectively. Furthermore, we analyze the effect of different marker set have on each of these classification tasks. With this large cohort of real clinical data, we demonstrate that our approach not only provides high hematological malignancies classification accuracy but further reveals insights on the discrminiability of the existing

[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]YFW, CCL, JLT are with Tai-Cheng Stem Cell Therapy Center, National Taiwan University, Taipei, Taiwan.

[3]BSK, JLT are with Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan.

[4]CCL is with Center of Stem Cell and Precision Medicine, Buddhist Tzu Chi General Hospital, Hualien, Taiwan.

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

TABLE I

THE DEMOGRAPHIC OF THE ENROLLED SAMPLES

Gender	Female	Male	NA	Total		
N	1299	1266	9	2574		
Age	<30	30-39	40-49	50-59	>=60	NA
N	440	471	436	506	697	24

marker panels used in the FC-based clinical diagnosis.

II. METHOD

A. Database and Study Population

The dataset used in this study is constructed retrospectively from the National Taiwan University Hospital. It consists of a total of 2574 specimen samples collected from 2009 to 2013. The demographics of the patients that conducted bone marrow aspiration is in Table I. After discarding incomplete data, 2424 specimen samples are used in our experiments (622 AML, 137 MDS, and 1665 normal). These enrolled bone marrow aspiration samples were examined using the flow cytometer (FASCalibur from Becton Dickinson Bioscience) with the marker set detailed in Table II. Each specimen FC data sample includes 11 tubes (each with a distinct pair of channel-antibodies) of 100,000 cells measured in 6 fluorescent channels (FSC, SSC, FITC, PE, PerCP, APC). All of the samples had previously been manual gated using "different-from-normal" approach by trained physicians to diagnose each speciman as one of the three mutually exclusive categories: "AML", "MDS", "normal". "AML" stands for newly diagnosed AML with residual leukemia cells exhibiting AML pattern. "MDS" is the newly diagnosed MDS with residual cell population exhibiting MDS pattern after treatment, and "normal" samples are specimens without abnormal cells. Our study is approved by the Research Ethic Committee of the National Taiwan University Hospital (No. 201705016RINA).

B. Cytometric Deep Phenotype Embedding

The overall framework is demonstrated in Figure 1. Our proposed cytometric deep phenotype embedding learning can be divided into two stages: a cell-level deep autoencoder followed by a specimen-level latent Fisher-scoring vectorization.

1) Cell-level Deep Autoencoder: The raw cytometry data of each tube is first transformed to a latent space using a per-tube autoencoder. Deep autoencoder (AE) is an well-known unsupervised deep network structure that is capable of learning to preserve information in the latent space through optimizing network weights with a reconstruction loss [9]. This AE-based latent space helps enhance representation power due to its capabilities of compactly represent the original data space's complex structure [10]. We use autoencoder with a 3-layer symmetric architecture (500,250,125 neurons and a 30 nodes latent layer). The activation function is selected to be relu. The deep autoencoder is implemented using the Keras (2015, GitHub) toolbox.

TABLE II

THE LIST OF MARKER-CHANNEL SETS.

	1st	2nd	3rd	4th	5th	6th
FITC	0	HLA-DR	CD5	CD56	CD16	CD15
PE	0	CD11b	CD19	CD38	CD13	CD34
PerCP	CD45	CD45	CD45	CD45	CD45	CD45
	7th	8th	9th	10th	11th	
FITC	CD14	CD7	CD2	HLA-DR	HLA-DR	
PE	CD33	CD56	CD117	CD34	CD117	
PerCP	CD45	CD45	CD45	CD45	CD45	
APC					CD34	

2) Specimen-level Latent Fisher-scoring Vectorization: In the second stage, we aim at representing each specimen with a single vector based on the transformed cell-level latent features, i.e., the output of the latent layer in the learned deep autoencoder. We use Fisher-scoring vectorization approach [11], i.e., an encoding approach combining both generative model with discriminative power. By assuming each tube's cell-level latent features are generative output of a multivariable Gaussian Mixture Model (GMM), we can summarize the cell characteristics (i.e., 100,000 cells) of each specimen sample by computing its Fisher-scoring function. A brief description is given below. We first pool all of the cell-level latent features of our dataset to learn a GMM distribution per tube. Then, let $X = x_t, t = 1...T$ be a set of T cells through the FC machine and p be a GMM probability distribution function (pdf) with parameters $\lambda = w_i, \mu_i, \Sigma_i, i = 1...K$, where w_i , μ_i and Σ_i are the weight, mean vector and covariance matrix for each mixture of Gaussian i. The gradient of log likelihood that characterizes each samples X can be derived by defining Fisher score function:

$$\nabla_\lambda \log p(X|\lambda)$$

where likelihood $p(x_t|\lambda) = \sum_{i=1}^{K} w_i p_i(x_t|\lambda)$ and thereby the posterior is given by the following:

$$\gamma_t(i) = p(i|x_t, \lambda) = \frac{w_i p_i(x_t|\lambda)}{\sum_{j=1}^{N} w_j p_j(x_t|\lambda)}$$

The gradient vectorization approach represents the direction for λ to better fit X with $p(X)$ by computing the first order and second order statistics,

$$g^X_{\mu_k} = \frac{1}{T\sqrt{w_k}} \sum_{i=1}^{T} r_t(k) \left(\frac{x_t - \mu_k}{\sigma_k} \right)$$

$$g^X_{\sigma_k} = \frac{1}{T\sqrt{2w_k}} \sum_{i=1}^{T} r_t(k) \left(\frac{(x_t - \mu_k)^2}{\sigma_k^2} - 1 \right)$$

the vector of $[g^X_{\mu_k}\ g^X_{\sigma_k}]$ is our proposed per-tube cytometric embedding output with 2*K*D dimensions where D is the feature dimension of raw cytometry data. These tube-wise specimen-level embeddings are L2-normalized and concatenated as the final input to the classifier. Our implementation is based on VLfeat and scikit-learn toolbox. The number of mixture set for GMM is obtained through grid search (specified as 16).

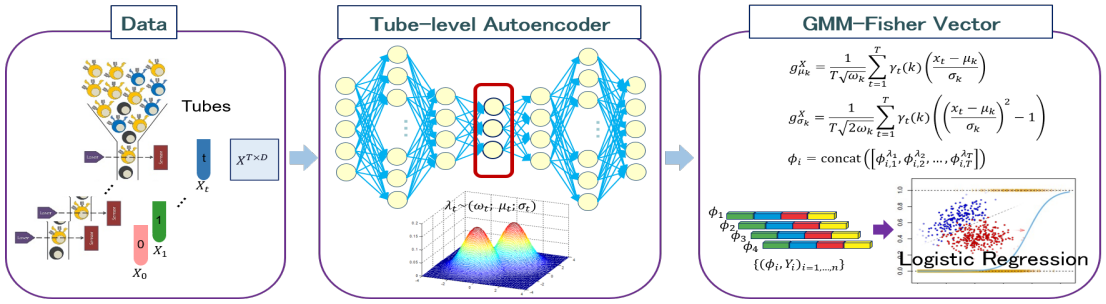


Fig. 1. The overall framework to learn deep cytometric phenotype embedding for AML and MDS classification.

C. Experimental Analyses

We use our proposed cytometric deep phenotype embedding as input to the logistic regression (LR) classifiers to perform four binary classification tasks: abnormal (AML+MDS) vs. normal, AML vs. normal, MDS vs. normal, and AML vs. MDS. In this study, we carry out 5-fold subject independent cross validation. In each fold, 20% of the entire database are left out for testing, and 80% of the database are used for training. To evaluate our model performances, we present results using accuracy (ACC), unweighted accuracy (UAR), and area under ROC curve (AUC). Two different analyses are carried out in this study: comparison to other specimen-level representations, and tube-selection experiments.

1) Comparison to other representations: We compare our method to different types of specimen-level representations listed below:

- Statistics Functional Encoding (SFE) method: We follow the same approach in the previous work [7], which uses 6 statistical functions for each raw cytometry data dimension to represent the 100,000 cells in a tube. The functions consist mean, standard deviation, skewness, kurtosis, median, and interquartile range.
- Posterior- based Phenotype Vector (P-P): This representation is derived by computing the average of cell-level posteriors of the targeted sample using the learned GMM model (similar method proposed in [8]).
- GMM-Fisher Phenotype Vector (GF-P): This representation directly applies GMM Fisher vectorization approach without the deep autoencoder cell-level learning.
- Deep Autoencoder based GMM-Fisher Phenotype Vector (AGF-P): Our proposed approach.

Our complete cytometric deep phenotype vector is a 10560 dimensional representation (960 dimension per tube).

2) Tube selection experiments: We conduct further analysis to identify the discriminability of the tube-wise marker-channel combination through tube-selection experiments for each of the four binary classification experiments. We firstly compute the average tube-wise ANOVA f-score for each dimension of our cytometric phenotype vector, and rank the importance of each tube according to this tube-wise average ANOVA f-score. We then train the LR by gradually incorporating each selected tube according to their importance values, i.e., the first model includes only the tube with the highest average f-score, and the second model adds the second highest-ranked tube, and so forth. Our aim is to identify the amount and the types of markers needed to reach the best classification performances.

III. RESULT

Table III reports the three accuracy metrics (accuracy, unweighted accuracy, and area under curve) for each of the four binary classification tasks (abnormal (AML+MDS) vs. normal, AML vs. normal, MDS vs. normal, and AML vs. MDS) obtained using different representations.

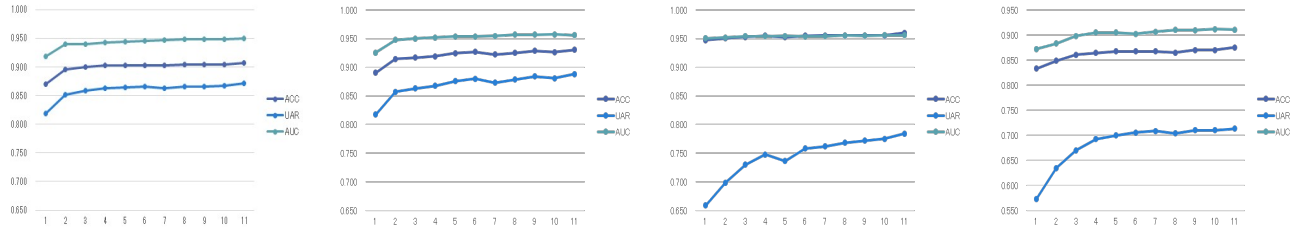
Generally, we observe that the AGF-P consistently outperforms other representations in all four of the tasks in terms of ACC, UAR and AUC. In task of abnormal vs. normal, AGF-P improves over SFE, P-P, GF-P, and AGF-P with 2.71%, 2.71%, and 1.06% relative improvement. It is worth noting that P-P does not perform competitively to GF-P, which demonstrates that expressive capability in the feature representation is crucial. By merely computing posterior probability as representation (P-P) is insufficient, which reflect in the lack of its discriminative power. On the other hand, SFE is comparable to GF-P in terms of UAR measures but not the ACC and AUC (a relative drop of 2.16% and 3.83% on average ACC and AUC across four tasks). The statistics functions are designed to capture the statistics properties of data distribution in the raw feature space directly, which may explain the reason that under certain circumstances, SFE would be comparable to GF-P since GF-P also attempts to characterize the raw feature space directly. However, the superior results of AGF-P when compared to SFE (2.27%, 2.69%, 4.81% relative UAR, ACC, AUC improvement) highlight the importance of cell-level feature representation power obtained through autoencoder.

The results of the tube selection experiments are shown in Figure 2. Interestingly, we observe that our method reaches around 90% on both ACC and AUC with 80% UAR by including only a single tube (the highest ranked in Table IV) for tasks of abnormal vs. normal and AML vs. normal. Furthermore, for tasks involving MDS, we see a general pattern that the accuracy obtained is lower as compared to other tasks, and they often require more tubes to perform well. This may related to the inherent disease categorization ambiguity of MDS and potentially the data imbalance issue. Nevertheless, an intriguing finding is that our method can achieve a similar high accuracy level using around 6 tubes as compared to the complete set of 11 tubes.

TABLE III

RESULTS OF THE FOUR DIFFERENT CLASSIFICATION TASKS USING DIFFERENT REPRESENTATIONS

	Abnormal vs Normal				AML vs Normal				MDS vs Normal				AML vs MDS			
LR	SFE	P-P	GF-P	AGF-P	SFE	P-P	GF-P	AGF-P	SFE	P-P	GF-P	AGF-P	SFE	P-P	GF-P	AGF-P
ACC	0.899	0.877	0.899	**0.906**	0.913	0.900	0.919	**0.931**	0.943	0.936	0.952	**0.960**	0.824	0.834	0.864	**0.875**
UAR	0.872	0.821	0.862	**0.872**	0.878	0.828	0.871	**0.888**	0.755	0.603	0.753	**0.784**	0.685	0.566	0.698	**0.713**
AUC	0.924	0.924	0.939	**0.949**	0.936	0.930	0.950	**0.956**	0.909	0.900	0.950	**0.956**	0.834	0.848	0.898	**0.911**



(a) Abnormal vs Normal (b) AML vs Normal (c) MDS vs Normal (d) AML vs MDS

Fig. 2. Results of tube selection are shown. The x-axis indicates number of tubes included with respect to the rank listed in Table IV.

IV. DISCUSSION

Our experimental results show an encouraging AUC of 94.9% (95% confidence interval [CI], 0.940-0.958) in classifying abnormal vs. normal specimens, and an average of 94.3% across four tasks. The slightly inferior results obtained in the AML and MDS is likely due to the the nature of the disease, where MDS is a continuum severity spectrum instead of a distinct disease by itself [12]. On the other hand, the tube selection results imply there exists a redundancy in the FC-based marker set chosen for diagnoses. This may also be attributed to the clinical practices of manual gating on 2D scatter plots, i.e., humans are not capable of visualizing directly in the high dimensional space, hence, certain level of redundancy is required. With proper computational methods, we imagine only a handful of tubes are required.

V. CONCLUSION

In this evaluation of applying cytometric deep phenotype embedding on modeling raw flow cytometry data, we demonstrate that our proposed method achieves highly accurate diagnostic performances in hematological malignancies classification tasks on the largest real patients FC sample cohort known to date. We will further pursuit the generalization of our approach to other hematologic diseases, e.g., acute lymphoblastic leukemia (ALL) or acute promyelocytic leukemia (APL). We hope to bring technologies in the status quo of current hematological diseases diagnostic practices through highly accurate and speedy assistive solutions.

TABLE IV

THE RANK OF TUBES IN DIFFERENT TASKS

rank	1	2	3	4	5	6	7	8	9	10	11
Abnormal vs. Normal	1	10	3	9	2	8	5	11	4	7	6
AML vs. Normal	1	10	3	8	5	9	2	11	7	4	6
MDS vs. Normal	9	4	10	11	2	3	1	5	8	7	6
AML vs. MDS	2	1	4	9	5	10	11	3	8	7	6

REFERENCES

[1] J. N. Saultz and R. Garzon, "Acute myeloid leukemia: a concise review," *Journal of clinical medicine*, vol. 5, no. 3, p. 33, 2016.

[2] F. Ravandi, R. B. Walter, and S. D. Freeman, "Evaluating measurable residual disease in acute myeloid leukemia," *Blood advances*, vol. 2, no. 11, pp. 1356–1366, 2018.

[3] C. Hung, W. Chen, P. Lai, C. Lin, and C. Lee, "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, July 2017, pp. 3110–3113.

[4] D. Wang, A. Khosla, R. Gargeya, H. Irshad, and A. H. Beck, "Deep learning for identifying metastatic breast cancer," *CoRR*, vol. abs/1606.05718, 2016.

[5] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus PhotographsAccuracy of a Deep Learning Algorithm for Detection of Diabetic RetinopathyAccuracy of a Deep Learning Algorithm for Detection of Diabetic Retinopathy," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 12 2016. [Online]. Available: https://dx.doi.org/10.1001/jama.2016.17216

[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 *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*. IEEE, 2015, pp. 438–441.

[7] M. Biehl, K. Bunte, and P. Schneider, "Analysis of flow cytometry data by matrix relevance learning vector quantization," *PLoS One*, vol. 8, no. 3, p. e59401, 2013.

[8] B. Rajwa, P. K. Wallace, E. A. Griffiths, and M. Dundar, "Automated assessment of disease progression in acute myeloid leukemia by probabilistic analysis of flow cytometry data," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 5, pp. 1089–1098, May 2017.

[9] C.-Y. Liou, W.-C. Cheng, J.-W. Liou, and D.-R. Liou, "Autoencoder for words," *Neurocomputing*, vol. 139, pp. 84–96, 2014.

[10] C. Song, F. Liu, Y. Huang, L. Wang, and T. Tan, "Auto-encoder based data clustering," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2013, pp. 117–124.

[11] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[12] J. C. Aster and R. M. Stone, "Clinical manifestations and diagnosis of the myelodysplastic syndromes," *UpToDate, Waltham, MA,[Acceso el 04 de Enero, 2016.]*, 2017.