Neighborhood-Correction Algorithm for Classification of Normal and Malignant Cells

Yongsheng Pan, Mingxia Liu, Yong Xia and Dinggang Shen

Abstract Classification of normal and malignant cells observed under a microscope is an essential and challenging step in the development of a cost-effective computer-aided diagnosis tool for acute lymphoblastic leukemia. In this paper, we propose the neighborhood-correction algorithm (NCA) to address this challenge, which consists of three major steps, including (1) fine-tuning a pretrained residual network using training data and producing initial labels and feature maps for test data, (2) constructing a Fisher vector for each cell image based on its feature maps, and (3) correcting the initial label of each test cell image via the weighted majority voting based on its most similar neighbors. We have evaluated this algorithm on the database provided by the grand challenge on the classification of normal and malignant cells (C-NMC) in B-ALL white blood cancer microscopic images. Experimental results demonstrate that our proposed NCA achieves the weighted F1-score of 92.50% and balanced accuracy of 91.73% in the preliminary testing and achieves weighted F1-score of 91.04% in the final testing, which ranks the first in C-NMC. Associated code is available at https://github.com/YongshengPan/ISBI-NMC.

Keywords Microscopic image classification · Fisher vector · Residual network · Leukemia · B-lymphoblast cells

Y. Pan · Y. Xia (✉)
Research & Development Institute of Northwestern Polytechnical University,
Shenzhen 518057, China
e-mail: yxia@nwpu.edu.cn

School of Computer Science and Engineering, Northwestern Polytechnical
University, Xi'an 710072, China

Y. Pan · M. Liu · D. Shen (✉)
Department of Radiology and BRIC, University of North Carolina,
Chapel Hill, NC 27599, USA
e-mail: dgshen@med.unc.edu

A. Gupta and R. Gupta (eds.), *ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging*, Lecture Notes in Bioengineering,
https://doi.org/10.1007/978-981-15-0798-4_8

1 Introduction

Acute lymphoblastic leukemia (ALL) is an acute cancer that affects millions of people globally, mostly children with age between 2 and 5 [15]. It is characterized by the development of plenty of immature lymphocytes in the bone marrow that interfere with the production of new blood cells. ALL is typically fatal within weeks or months if with less treatment and there is no reliable measure to prevent it. Thus, it is critical to perform early diagnosis to improve the survival rate.

Diagnosis of ALL is typically based on blood tests and bone marrow examination [5]. The mechanism involves that the genetic mutations result in rapid cell division and increased immature lymphocytes [10]. In advanced cancer stages, the cancer cells start growing in an unrestricted fashion, and therefore they are presented in much larger numbers as compared to the numbers in an early-stage patient. This could be identified by advanced methods such as flow cytometry [17], which, however, are too expensive to be widely used in less-developed regions.

As the morphology of cells can be observed under a microscope, it is possible to develop a cost-effective computer-aided solution to normal–malignant B-lymphoblast cell classification on microscopic images and hence to reduce the cost of ALL diagnosis. Recent years have witnessed the incredible advances in automated image classification [11, 12], including aggregative models [16] and deep convolutional neural networks (DCNNs) [12]. Since the image representation ability of DCNNs learned from large-scale datasets can be transferred to extract local descriptors on a generic image classification task, DCNNs have been combined with FV to boost the performance of image classification [1, 13, 14]. The Fisher vector (FV) and residual network (ResNet) [8] are typical representatives of aggregative models and DCNNs, respectively. The FV treats each image as a set of local visual descriptors sampled from a distribution under the label condition, and thus aggregates these descriptors into high-level image representations that can infer semantics and be separated linearly. The ResNet provides a uniform framework of feature extraction and classification to free users from the troublesome handcrafted feature extraction.

Despite these advances, it still remains a challenge to differentiate malignant cells from normal ones accurately since the morphological specificity of malignant cells appears very similar to that of normal cells [3, 4]. The grand challenge on classification of normal versus malignant cells (C-NMC) in B-ALL white blood cancer microscopic images organized in conjunction with the 2019 IEEE international symposium on biomedical imaging (ISBI 2019) recognized the significance and complexity of this image classification task and provided a database of clinical cells to evaluate various image classification approaches [7].

In this paper, we propose the neighborhood-correction algorithm (NCA) for the classification of normal–malignant B-lymphoblast cell images. To combine the advantages of ResNet and FV, we first fine-tune a pretrained ResNet (pRN) using training data and then use feature maps of each test image produced by the fine-tuned ResNet (fRN) as local descriptors to construct FVs. Finally, we correct the initial

Table 1 Distribution of C-NMC Challenge database

		Phase-I			Phase-II	Phase-III
		Fold 1	Fold 2	Fold 3		
Subjects	Cancer	19	11	17	13	9
	Normal	9	3	14	15	8
Cell images	Malignant	2397	2418	2457	1219	–
	Normal	1130	1163	1096	648	–

label predicted by fRN of each test cell image using the weighted majority voting based on its most similar neighbors. Our algorithm has been evaluated on the C-NMC Challenge database and achieved state-of-the-art performance.

2 Materials

The C-NMC Challenge database [7] contains the cell images from 84 cancer and 70 normal subjects, where the label of each cell is assumed to be the subject-level label. This dataset was released in three phases. 7272 malignant cells from 47 cancer subjects and 3389 normal cells from 26 normal subjects, together with a 3-fold data split, were released in Phase-I. 1219 malignant cells from 13 cancer subjects and 648 normal cells from 15 normal subjects were released in Phase-II. In Phase-III, the 2586 cell images from 9 cancer subjects and 8 normal subjects were released, but the label of each cell was withheld for online validation. The data distribution is illustrated in Table 1. All cells have been preprocessed via stain normalization and cell segmentation [2–4, 6, 7]. Each cell is roughly of size 300 × 300 and relocated at the center of a 450 × 450 baseplate.

3 Method

The proposed NCA mainly contains three steps. In Step 1, a pRN is fine-tuned using the training data to get the fRN, which can predict the initial labels and feature maps for test cells. In Step 2, the feature maps of each cell are aggregated to an FV, followed by calculating the dot product of FVs to measure the similarity of each pair of cells. In Step 3, the initial label of each cell is adjusted according to its neighborhood in the FV space. A diagram that summarizes our proposed model is shown in Fig. 1.

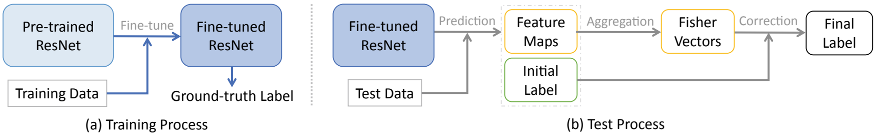

Fig. 1 Diagram of the proposed NCA algorithm for normal–malignant cell classification on microscopic images

3.1 Fine-Tuning ResNets

Since the C-NMC database is too small to train a DCNN with random initialization, we adopt the pRN, whose three off-the-shelf models (e.g., pRN50, pRN101, and pRN152) pretrained on the ImageNet database are available at the MatConvNet website [19], to avoid falling into overfitting when training. The architecture of a pRN contains a general convolutional layer followed with a max-pooling layer, dozens of residual blocks, and a fully connected layer. To apply the pRN to the normal–malignant cell classification, we keep only the first two neurons of the last fully connected layer and remove other neurons and related weights in this layer. While fine-tuning the pRN, each cell image is randomly rotated 0–350° for data augmentation and cropped to 224 × 224 to fit the input size in each of 20 epochs. The stochastic gradient descent (SGD) algorithm [18] is used as the optimizer with a batch size of 32 and a learning rate of 0.001 for the first 10 epochs and 0.0001 for the next 10 epochs. After obtaining the fRN, we rotate each test cell image 0–350° with an interval of 10°, thus resulting in 36 augmented copies, and then feed the cropped copies (size: 224 × 224) to fRN. The initial label of each test image is the majority voting result of its 36 augmented copies.

3.2 Combined ResNet-FV for Cell Representation

We extract the $7 \times 7 \times 512$ feature maps produced by the "res5c_branch2a" layer of the fRN as local descriptors. Since each cell image has 36 augmented copies, we totally extract $N = 36 \times 7 \times 7$ descriptors, denoted as $X = \left\{x_n \in R^D; n = 1, \ldots, N\right\}$ $(D = 512)$. We assume that these descriptors of all cell images follow a Gaussian mixture model (GMM) u_θ with diagonal covariance, whose parameters $\theta = \{\omega_l, \mu_l, \sigma_l; l = 1, \ldots, L\}$ are estimated by the maximum likelihood estimation (MLE). Then, the FV of X is defined as the Fisher information normalized gradient of the log likelihood [7] of X with respect to each mean μ_l and standard deviation σ_l, shown as follows:

$$\mathcal{B}_\theta = (\mathcal{B}_{\mu,1}{}^{\mathrm{T}}, \ldots, \mathcal{B}_{\mu,L}{}^{\mathrm{T}}, \mathcal{B}_{\sigma,1}{}^{\mathrm{T}}, \ldots, \mathcal{B}_{\sigma,L}{}^{\mathrm{T}})^{\mathrm{T}}, \tag{1}$$

where

$$\begin{cases} \mathcal{B}_{l,\mu} = \frac{1}{N\sqrt{\omega_l}} \sum_{n=1}^{N} \tau_l(x_n) \left[\frac{x_n - \mu_l}{\sigma_l} \right], \\ \mathcal{B}_{l,\sigma} = \frac{1}{N\sqrt{2\omega_l}} \sum_{n=1}^{N} \tau_l(x_n) \left[\frac{(x_n - \mu_l)^2}{\sigma_l^2} - 1 \right], \end{cases} \tag{2}$$

and $\tau_l(x_n) = \frac{\omega_l \mu_l(x_n)}{\sum_{j=1}^{L} [\omega_j u_j(x_n)]}$ is the posterior probability of x_n being assigned to the lth Gaussian component u_l. To cancel out the effect that cells at different grades of maturity may have different sizes and thus result in biased FV signatures, the following l_2-normalization [16] is then employed

$$\mathfrak{B}_\theta^X = \mathcal{B}_\theta^X / \left\| \mathcal{B}_\theta^X \right\|_2 \tag{3}$$

The Fisher kernel (FK) that measures the similarity between two cell images X_1 and X_2 is defined as

$$J_{FK}(X_1, X_2) = \mathfrak{B}_\theta^{X_1\,\mathrm{T}} \mathfrak{B}_\theta^{X_2} \tag{4}$$

Incidentally, this kernel can be cooperated with a support vector machine (SVM) [1] for cell image classification.

3.3 Label Correction

After obtaining the FV and initial label of each testing cell image, we can use the neighborhood information to correct misclassified images. We assume that the images with similar discriminative representations (i.e., FVs) belong to the same category. Suppose that $X_{i,(1)}, X_{i,(2)}, \ldots, X_{i,(K)}$ are the most similar K images of X_i, whose initial labels are $y_{i,(1)}, y_{i,(2)}, \ldots, y_{i,(K)}$, respectively, where $X_{i,(1)} = X_i$ and $y_{i,(1)} = y_i$ since the most similar image is itself. If the label of X_i is different from any of its K nearest neighbors in the FV space, we adjust the label of X_i as the weighted voting result of $y_{i,(1)}, \ldots, y_{i,(K)}$ as follows:

$$y_i^* = \mathrm{sign} \left[\sum_{k=1}^{K} J(X_i, X_{i,(k)}) y_{i,(K)} \right] \tag{5}$$

where $J(X_1, X_2)$ equals $J_{FK}(X_1, X_2)$ if only one fRN is used, or the minimum of J_{FK} values if multiple fRNs are used.

4 Experiments and Discussions

We performed three groups of experiments to evaluate our NCA on the C-NMC database. We first evaluated the impact of different ResNets, and then compared NCA to three methods on the Phase-I dataset. We finally applied NCA to the Phase-II dataset. We set $L = 8$ in the aggregative model and $K = 7$ for label correction. The accuracy of each method is measured by the weighted F1-score (WF1S) and the balanced accuracy (BACC) [9].

4.1 Cross-Validation of Baseline Method

We first evaluated the baseline method (BM) that directly applied fRN to test data via 3-fold cross-validation. Each time, one of the three folds was used to test the fRN fine-tuned on the other two folds. The WF1S and BACC values achieved by BM (including fRN50, fRN101, and fRN152) were shown in Fig. 2, where the horizontal axis indicates the performance on each of the three folds and the average (AVG) performance. Figure 2 reveals that (1) fRN101 results in slightly higher average WF1S and BACC values than fRN50 and fRN152, (2) no matter which fRN was used, using Fold-0 and Fold-2 as the test dataset leads to the highest and lowest accuracy, respectively, largely due to the inconsistency over subjects, and (3) the performance variation caused by different fRNs is much less than that caused by different test datasets.

4.2 Performance Gain Caused by Label Correction

In the second group of experiments, we evaluated the effectiveness of our NCA by comparing it to BM and FV with the linear SVM as the classifier. Herein, Res-Net-50 (with 50 layers) was used for all methods. For FV, we considered both the FV with

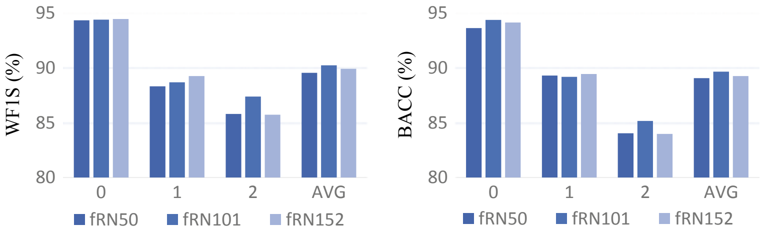

Fig. 2 WF1S (left) and BACC (right) values (%) obtained by fRN50/101/152 in 3-fold cross-validation. The horizontal axis indicates the result on Fold-0/1/2 and the average result

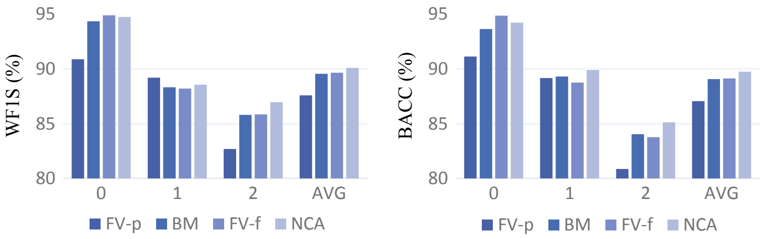

Fig. 3 WF1S (left) and BACC (right) values (%) obtained by four methods based on ResNet-50, where NCA produces the highest metrics on average (AVG)

the pRN50 (FV-p) and the FV with fRN50 (FV-f). The WF1S and BACC values for three folds and their average are shown in Fig. 3. From Fig. 3, we can see that NCA generally outperforms three competing methods regarding the average values of both metrics. Meanwhile, FV-f and NCA perform the best on Fold-0 and Fold-2, respectively while NCA and FV-p reach the best WF1S and BACC on Fold-1, respectively. Compared to BM, the improved BACC/WF1S values of NCA are 0.59/0.39%, 0.59/0.24%, and 1.09/1.15% for Fold-0, Fold-1, and Fold-2, respectively. All these results indicate that our NCA is successful in C-NMC and is useful for ALL diagnosis. Besides, BM, FV-f, and NCA are superior to FV-p obviously while FV-f reaches a comparable but slightly better performance than BM. It indicates that the fine-tuned pRN model can increase the distinguishability for cell representation and the FK is good at measuring the similarity of cells. Therefore, it is reasonable to use FV with fRN to estimate the data distribution of cell images. Particularly, our NCA achieves the largest improvement on Fold-2, on which the other three methods achieve the worst metrics. This suggests that our NCA may have an outstanding capability to boost the unsatisfying performance.

4.3 Results on Phase-II Dataset

In the third group of experiments, we used the Phase-II dataset to further evaluate our NCA against BM. We considered different combinations that use one, two, or three folds in Phase-I for training, with Phase-II used for external validation. The BACC and WF1S values are reported in Table 2. While using single fold for training, the models trained on Fold-1 perform much worse than those trained on Fold-0 or Fold-2, almost no better than random guesses. The reason lies in the fact that Fold-1 contains only three normal subjects, which results in a large negative bias to the trained models. While using two folds for training, the models trained on Fold-0 and Fold-1 perform much worse than those on Fold-0 and Fold-2 and those on Fold-1 and Fold-2. This again verifies that the cells among different folds (or different subjects) are inconsistent, which is attributed mainly to the heterogeneous data acquisition process. Also, the models trained on all three folds perform obviously better than the

Table 2 WF1S/BACC of 21 pairs of comparisons on Phase-II (%)

Fold			Method	ResNet-50		ResNet-101		ResNet-152	
0	1	2		WF1S	BACC	WF1S	BACC	WF1S	BACC
✓			BM	78.54	78.23	78.78	79.01	77.37	76.68
			NCA	**81.42**	**81.89**	**82.87**	**83.77**	**79.25**	**79.07**
	✓		BM	**52.88**	**56.46**	**48.52**	**53.91**	**52.53**	**56.00**
			NCA	50.62	55.24	44.72	52.46	49.08	54.13
		✓	BM	76.70	78.95	80.32	82.28	76.65	78.39
			NCA	**77.40**	**80.22**	**80.98**	**83.27**	**77.67**	**80.33**
✓	✓		BM	67.82	66.80	70.30	68.92	72.46	70.98
			NCA	**72.90**	**71.32**	**73.81**	**72.00**	**76.00**	**74.19**
✓		✓	BM	83.55	84.41	83.55	84.27	80.60	81.27
			NCA	**84.32**	**85.93**	**84.21**	**85.67**	**82.04**	**83.39**
	✓	✓	BM	84.01	83.35	82.79	82.34	81.32	80.27
			NCA	**88.11**	**87.89**	**84.37**	**84.19**	**85.41**	**84.38**
✓	✓	✓	BM	83.69	83.00	84.27	84.11	84.11	83.64
			NCA	**88.41**	**87.99**	**85.51**	**85.92**	**85.41**	**85.88**

average of those models trained on one or two folds and are even similar to the best models trained on two folds. This indicates that using more data will further improve the performance.

Except for the results on only Fold-1, our NCA consistently outperforms BM on both metrics; even the increments are different on different fold combinations and different ResNets. When all three folds are used, the average results of BM and NCA over three ResNets are 84.02/83.58% and 86.44/86.60%, respectively, achieving an increment of 2.42/3.02%. Particularly, NCA achieves the best result of 88.41/87.99% on ResNet-50. Using the ensemble of fRN50, fRN101, and fRN152, NCA achieves 92.50% and 91.73% of WF1S and BACC, respectively, which achieves the top result on the leaderboard of preliminary testing. NCA also ranks the first (91.04% of WF1S) in the final testing (Phase-III) while using data in both Phase-I and Phase-II to fine-tune ResNets.

4.4 Feedback to BM

In the third group of experiments, we have corrected the initial labels of cell images in Phase-II with obtaining fRNs on Phase-I and those in Phase-III with obtaining fRNs on Phase-I and Phase-II. Actually, these cell images with corrected labels can be further cooperated with these cell images with real labels to improve the BM. For example, we can improve the BM by fine-tuning the pRN-50 with both the cells with a real label in Phase-I and the cells with a corrected label in Phase-II. The performance

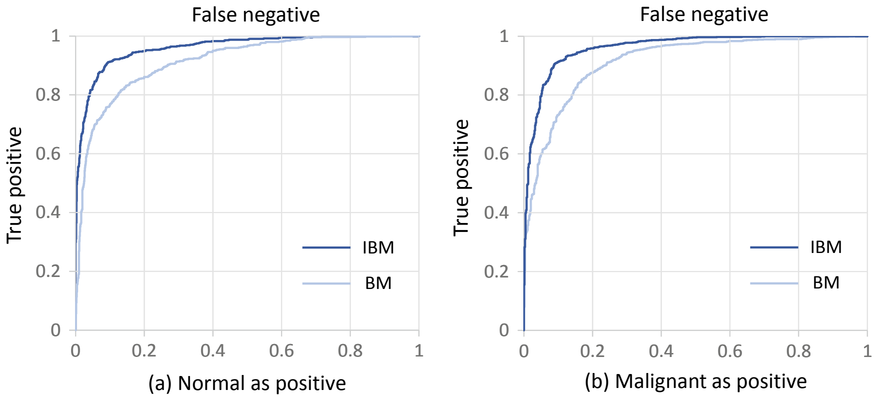

Fig. 4 WF1S (left) and BACC (right) values (%) obtained by four methods based on ResNet-50, where NCA produces the highest metrics on average (AVG)

curves of BM and the improved BM (IBM) are illustrated in Fig. 4, where Fig. 4a is the case considering normal cells as positive and Fig. 4b is the case considering malignant cells as positive. It can be seen that in both cases, IBM outperforms BM obviously. The area under the curve (AUC) is 91.52%, and 96.14% for BM and IBM, respectively, which indicates a huge improvement of 4.62%.

5 Conclusion

We have developed a neighborhood-correction algorithm (NCA) for normal–malignant cell classification on microscopic images. Results on the C-NMC database show that the NCA achieved acceptable accuracy on cell image classification. Our future work will focus on developing a unified framework for feature learning and label correction.

Acknowledgements Y. Pan and Y. Xia were supported in part by the Science and Technology Innovation Committee of Shenzhen Municipality, China, under Grants JCYJ20180306171334997, in part by the National Natural Science Foundation of China under Grants 61771397, in part by the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University under Grants CX201835, and in part by the Project for Graduate Innovation team of Northwestern Polytechnical University.

References

1. Cimpoi, M., Maji, S., Kokkinos, I., Vedaldi, A.: Deep filter banks for texture recognition, description, and segmentation. Int. J. Comput. Vis. **118**(1), 65–94 (2016)
2. Duggal, R., Gupta, A., Gupta, R.: Segmentation of overlapping/touching white blood cell nuclei using artificial neural networks. In: CME Series on Hemato-Oncopathology. All India Institute of Medical Sciences (AIIMS) (2016)
3. Duggal, R., Gupta, A., Gupta, R., Mallick, P.: SD-layer: stain deconvolutional layer for CNNs in medical microscopic imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 435–443. Springer (2017)
4. Duggal, R., Gupta, A., Gupta, R., Wadhwa, M., Ahuja, C.: Overlapping cell nuclei segmentation in microscopic images using deep belief networks. In: Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing, p. 82. ACM (2016)
5. Ferri, F.F.: Ferri's Clinical Advisor 2018 E-Book: 5 Books in 1. Elsevier Health Sciences (2017)
6. Gupta, A., Duggal, R., Gupta, R., Kumar, L., Thakkar, N., Satpathy, D.: GCTI-SN: geometry-inspired chemical and tissue invariant stain normalization of microscopic medical images (under review)
7. Gupta, R., Mallick, P., Duggal, R., Gupta, A., Sharma, O.: Stain color normalization and segmentation of plasma cells in microscopic images as a prelude to development of computer assisted automated disease diagnostic tool in multiple myeloma. Clin. Lymphoma Myeloma Leukemia **17**(1), e99 (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
9. Hripcsak, G., Rothschild, A.S.: Agreement, the F-measure, and reliability in information retrieval. J. Am. Med. Inform. Assoc. **12**(3), 296–298 (2005)
10. Hunger, S.P., Mullighan, C.G.: Acute lymphoblastic leukemia in children. New Engl. J. Med. **373**(16), 1541–1552 (2015)
11. Li, Z., Song, Y., Mcloughlin, I., Dai, L.: Compact convolutional neural network transfer learning for small-scale image classification. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2737–2741. IEEE (2016)
12. Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear convolutional neural networks for fine-grained visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1309–1322 (2018)
13. Liu, L., Wang, P., Shen, C., Wang, L., Van Den Hengel, A., Wang, C., Shen, H.T.: Compositional model based Fisher vector coding for image classification. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2335–2348 (2017)
14. Pan, Y., Xia, Y., Shen, D.: Foreground Fisher Vector: Encoding Class-Relevant Foreground to Improve Image Classification. IEEE (accepted on 2019)
15. Pui, C.H.: Acute Lymphoblastic Leukemia. Springer (2011)
16. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the Fisher vector: theory and practice. Int. J. Comput. Vis. **105**(3), 222–245 (2013)
17. Shapiro, H.M.: Practical Flow Cytometry. Wiley (2005)
18. Teng, M., Wood, F.: Bayesian distributed stochastic gradient descent. In: Advances in Neural Information Processing Systems, pp. 6380–6390 (2018)
19. Vedaldi, A., Lenc, K.: Matconvnet: convolutional neural networks for MATLAB. In: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 689–692. ACM (2015)