Leukemia Blood Cell Image Classification Using Convolutional Neural Network

T. T. P. Thanh, Caleb Vununu, Sukhrob Atoev, Suk-Hwan Lee, and Ki-Ryong Kwon

***Abstract*—Acute myeloid leukemia is a type of malignant blood cell cancer that can affect both children and adults. There are 60,140 people were expected to be diagnosed with Leukemia in 2016, according to the Leukemia and Lymphoma Society. In order to get the most effective treatment, the patient needs early diagnosis. Therefore we need to have a support system of early diagnosis to guide treatment for patients with acute leukemia as soon as possible. In this paper, the authors propose a Convolutional Neural Network (CNN) based method to distinguish normal and abnormal blood cell images. The proposed method achieves an accuracy up to 96.6% with the dataset including 1188 blood cell images.**

***Index Terms*—Classification, convolutional neural network, leucocyte, leukemia.**

I. Introduction

Leukemia (blood cancer) is a cancer of blood cells caused by radiation exposure, family history of leukemia and exposure to certain chemicals [1].

In general, leukemia was classified based on the speed of progression and the type of cells. Base on leukemia progresses, the first type of Leukemia classification is divided into two groups: acute leukemia and chronic leukemia. In acute leukemia, the abnormal blood cells (immature blood cells) which cannot carry out their normal functions are multiply speedily. In chronic leukemia, some types of it produce too many cells and some cause too few cells are born. In contrast to acute leukemia, chronic leukemia concern mature blood cells. The second type of leukemia, which is determined by the type of white blood cell affected, consists of lymphocytic leukemia and myelogenous leukemia. Lymphocytic leukemia (lymphoblastic) occurs in a type of marrow cell that forms lymphocytes. Myelogenous leukemia (myeloid) affects myeloid cells that give rise to red blood cells, some other types of white cells and platelets.

Combining these two general classifications above, leukemia was classified into four main types based on severity level and infected cells type - acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), chronic lymphocytic leukemia (CLL) and chronic myeloid leukemia (CML), as we see in Table I [2].

Acute lymphoblastic leukemia is not only the most common type of Leukemia in young children, but also affects adults in the age of 65 and above 65 years old.

Manuscript received January 9, 2018; revised April 3, 2018.

T. T. P. Thanh, Caleb Vununu, Sukhrob Atoev, and Ki-Ryong Kwon are with Department of IT Convergence and Applications Eng., Pukyong National University, the Republic of Korea (e-mail: thanhttp02@gmail.com, exen.xmen@gmail.com, sukhrob.reus@mail.ru, krkwon@pknu.ac.kr).

Suk-Hwan Lee is with Department of Information Security, Tongmyong University, the Republic of Korea (e-mail: skylee@tu.ac.kr).

Acute myeloid leukemia occurs more commonly in adults than in children and more commonly in men than women. AML is listed as the most dangerous type of Leukemia because there is only 26.9% surviving rate over the five-year period [3].

Chronic lymphocytic leukemia is more common at the age of 55 and older and it occurs mainly in men with two-thirds of patients are men. The five-year survival rate of CLL in the 2007-2013 period is 83.2% [4].

Chronic myeloid leukemia occurs mainly in adult with the five-year survival rate is 66.9%

According to the National Cancer Institute, it is estimated that there are 24,500 people died because of leukemia in the US in 2017. Leukemia represents 4.1% of all cancer cases deaths in the U.S. [5].

Acute leukemia diagnosis is used to base on a bone marrow examination by a pathologist, the test result is based on the experience of the technician. Therefore, an automatic system to early diagnosis leukemia has an important role in Leukemia diagnostic system.

TABLE I: Four Main Types of Leukemia

	Lymphocytic leukemia	Myelogenous leukemia
Acute	Acute lymphoblastic leukemia (ALL)	Acute myeloid leukemia (AML)
Chronic	Chronic lymphocytic leukemia (CLL)	Chronic myeloid leukemia (CML)

II. Related Work

Researches about leukemia classification in recent years are mainly based on computer vision techniques [6], [7]. The most common algorithm in this approach consists of several rigid steps: image pre-processing, clustering, morphological filtering, segmentation, feature selection or extraction, classification, and evaluation [8]. Most of the authors in the literature have adopted machine learning techniques such as K-means clustering in order to detect and classify blood cells in images. In most of cases, the conventional statistical features such as energy, entropy, contrast, and correlation, were extracted and given as inputs to a machine learning model. It is clear that the traditional machine learning methods have some disadvantages such as time-consuming in development and, mostly, the need of deciding which kind of features must be utilized in order to maximize the classification's accuracy. Instead, deep learning can learn and extract high level features automatically and perform classification in the same time. Therefore, we propose a novel Convolutional Neural Network (CNN) architecture to discriminate normal and abnormal blood cell images. The

DOI: 10.7763/IJCTE.2018.V10.1198

advantage of using CNN is not only it reduces the processing time by allowing us to skip most of the pre-processing steps, but also it has the ability of extracting features that are better than the conventional statistical features, as we will demonstrate in this paper.

III. Proposed Method

A. Proposed CNN Architecture

In this paper, we use CNN to extract features from raw blood cell images and perform classification. The architecture of CNN includes three main types of layers: convolutional layer, pooling layer, and fully-connected layer. Convolution layers compute the output of neurons by calculating a weighted sum of the inputs, adding a bias to that weighted sum and then applying the rectifier linear unit (ReLu) on it. ReLu, given by the equation ReLu (z) = max (0, z), is one type of activation function which decides whether a neuron should be active or not. Pooling layers are in charge of reducing the spatial size of the representation in order to decrease the number of parameters and computations, leading to control overfitting. Fully connected layers contain neurons that are connected to all the activations from the previous layer.

In this work, we use a network containing 7 layers, as we see in Fig. 1. The first 5 layers perform feature extraction and the other 2 layers (fully connected and softmax) classify the extracted features. The input image has the size of 100x100x3. In the convolution layer 1, we used a filter size of 5x5 and a total of 16 different filters. The stride is 1 and no zero-padding was applied.

The second and third convolution layers have the same structure with the first one but different number of filters, 32 and 64, respectively. We used pooling layer with filter size 2, stride 2 to decrease the volume spatially. During the learning, the chosen size of the mini batch was 100. ReLu is used as the activation function.

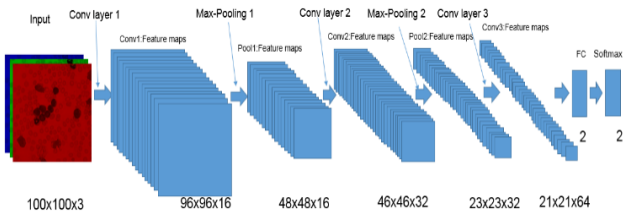


Fig. 1. The architecture of the network.

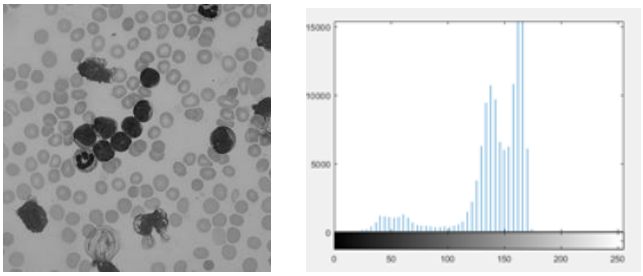
In the preliminary version of this work that has been published in the 6th International Conference on Advanced Information Technologies and Applications (ICAITA 2017) [9], the authors have also adopted the convolutional neural network but with a slightly shallow architecture compared to the one proposed here and a really small number of data. In the current research, we present a deeper architecture trained on a significantly augmented dataset. Details about data augmentation are discussed in the next section.

B. Data Augmentation

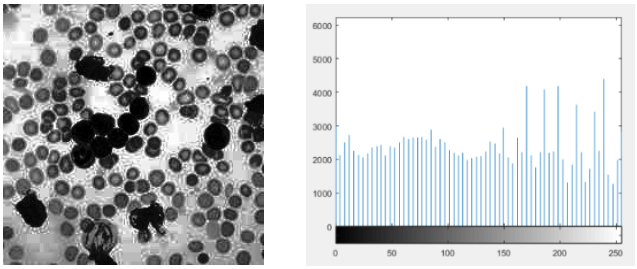
The main problem that bioinformatics researchers face when finding solutions for detecting and diagnosing Leukemia diseases is a lack of dataset because medical images are private. Besides, the more Leukemia images CNN can handle, the higher accuracy achieved. Therefore, the need for a large enough dataset to build an effective CNN architecture in the diagnosis of Leukemia is extremely urgent.

In this paper, the authors propose some simple data augmentation methods to expand the available data. [10], [11].

1) Histogram equalization



Grayscale ALL image | Grayscale ALL histogram



Equalized image | Equalized Histogram

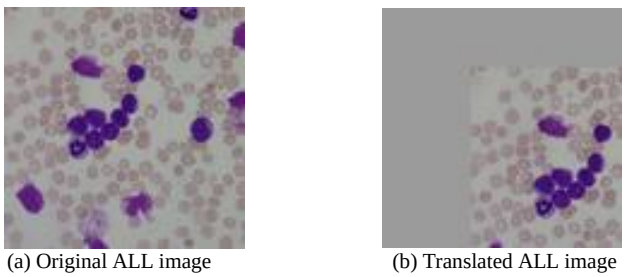
Fig. 2. The grayscale ALL image and the adjusted image.

Histogram equalization is a simple technique to enhance the contrast of a low contrast image by detecting the distribution of pixel densities and stretching the range of intensity values to a desired range of values. [12]

As in Fig. 2, the equalized image show the more detail of white blood cells and red blood cells in ALL image.

2) Translation

In this paper, the authors perform a translation operation to shift an image along both x-axis and y-axis with the displacement value between 25 and the middle of each axis.



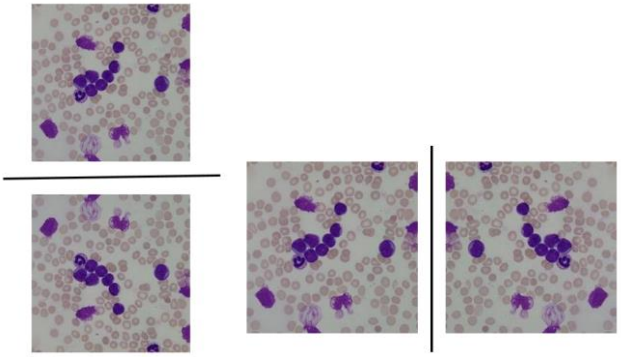
(a) Original ALL image | (b) Translated ALL image

Fig. 3. The original image and translated image.

For example, Fig. 3 shows the original ALL image and the translated ALL image with the displacement belong x-axis is 84 and belong y-axis is 56 on the left side and right side, respectively.

3) Reflection

Image reflection is a method to mirror an image through the vertical and horizontal axis as shown in Fig. 4. This is a simple solution to extend dataset but it proves to high efficiency because the mirrored image still remains the content of the image.

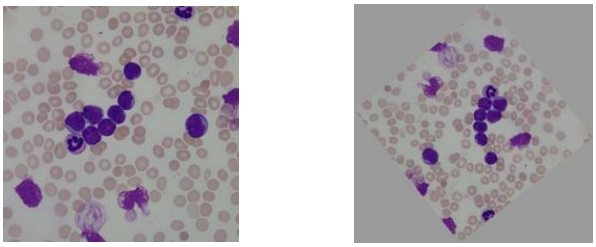


(a) Horizontal axis (b) Vertical axis

Fig. 4. Image reflection through vertical and horizontal axis.

4) Rotation

We handle rotations of images by randomly rotating images right or left from our original dataset with the rotation range between [-180 °, 180 °] as illustrated in Fig. 5.



(a) Original ALL image (b) Rotation image

Fig. 5. Original ALL image and rotation image.

5) Shearing image

Image is sheared along x and y-axis by mapping a pair of input coordinates [x, y] to a pair of output coordinates $[x', y']$ [13] as in Fig. 6.

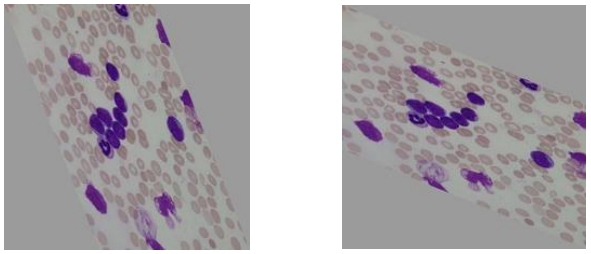
Shearing along x-axis as bellow:

$$x' = x + ay$$
$$y' = y$$
$$[x' \quad y' \quad 1] = [x \quad y \quad 1] * \begin{bmatrix} 1 & a & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Shearing along y-axis:

$$x' = x$$
$$y' = bx + y$$
$$[x' \quad y' \quad 1] = [x \quad y \quad 1] * \begin{bmatrix} 1 & 0 & 0 \\ b & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

In the formulas above, a and b are generated randomly between 0.3 and 1.



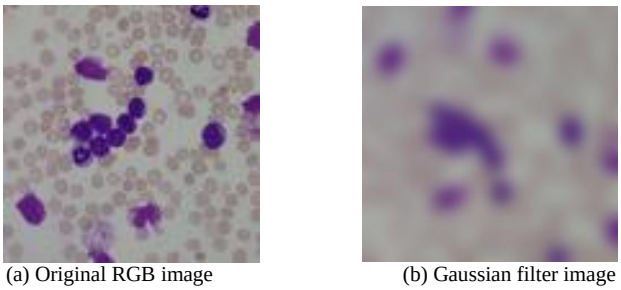
(a) Sheared image along x-axis (b) Sheared image along y-axis

Fig. 6. The original ALL image and sheared image.

6) Grayscale image

We convert all of the images in ALL-IDB1 dataset from RGB format to grayscale image. This is a simple method but it has applicability to simply creating more data.

7) Blurring image (blurring RGB and grayscale image)



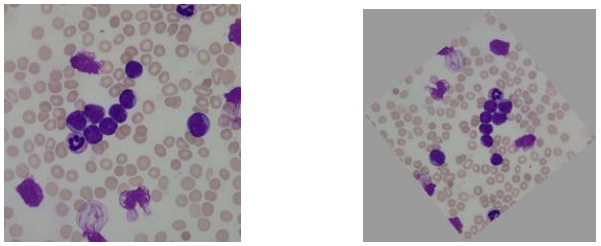
(a) Original RGB image (b) Gaussian filter image

Fig. 7. The original ALL image and Gaussian filter image.

Fig. 7 shows how to perform blurring operation in original ALL image with standard deviation is 5.11.

In this work, we apply a Gaussian filter to blur RGB image and grayscale image to generate 98 blurring images. The standard deviation for the Gaussian filter in our experiment is set up randomly between [4], [8].

8) Rotation image



(a) Original image (b) Rotated image

Fig. 8. The original image and rotated image.

Original image is rotated a rotation angle between $[-180^o, 180^o]$ randomly. If the rotation angle is a positive number, the image rotates counterclockwise. Otherwise, if the rotation angle is a negative number, the image rotates clockwise.

The output rotated image must be large enough to contain the original image. Therefore, the output image has bigger size than the input image and pixels that fall outside the boundary of the original image is set to 155. After rotating the image, resize the output image to be the same size with the input image.

IV. Experimental Result

In [9], the original ALL-IDB1 image database which consists of 108 cell image (59 normal and 49 abnormal cell images) was used. In this paper, to increase the accuracy of our proposed method, we have augmented the original dataset to 1188 pictures by applying transformations such as blurring, histogram equalization, reflection, translation, rotation, shearing. Blurring use Gaussian filter with sigma σ generated randomly with a uniform distribution between [4], [8]. Image reflection was applied along the horizontal and vertical axis, image rotation towards right and left with an angle randomly picked between -180 ° and 180 °. We have also applied horizontal and vertical translation for images with the displacement value chosen randomly with a uniform distribution between 25 and the middle of the axis. With shearing transformation, horizontal and vertical shear were used with the value of distortion randomly generated between [0.3, 1].

TABLE II: CNN CLASSIFIER RESULT

	Testing Phase		
	Test Set	Correct Classification	Misclassification
Normal cell	162	152	10
Abnormal cell	195	193	2
Total	357	344	13

Our experiments were conducted on Matlab with 1188 images, 70% (831 images) of them for training and the remaining 30% (357 images) for testing our model. The slightly narrow architecture used in [9] dramatically failed to reach an appropriate accuracy when applied to this augmented dataset. Therefore, we have presented here a deeper CNN architecture and changed the size of the input volume in order to improve the accuracy rate of the recognition of leukemia (our proposed CNN model achieved 96.6%). In Table II, our proposed CNN architecture predicted that 152 images were normal cell images, of the 162 actual normal cell images, and of the 195 abnormal cell images, it predicted that 193 images were abnormal cell images and 2 were normal cell images.

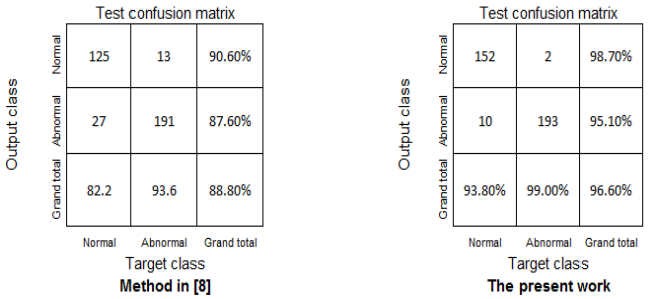


Fig. 9. The accuracy rate for the method used in [8] and the accuracy of the present work.

For the purpose of evaluation, we have compared our method with the conventional statistical features used in [8]. We remind that the experiments in [8] were conducted on a really small dataset. Once applied on a significantly augmented dataset, the conventional statistical features are outperformed by the CNN generated features. In Fig. 2, we can see that the accuracy rate after using our CNN model is significantly better than when utilizing the conventional statistical features as used in [8].

V. CONCLUSION

This method of leucocyte classification promises to be used in diagnostic systems for leukemia for early detection of disease in the daily life. The authors have performed the proposed method in a largely augmented dataset in order to confirm the accuracy and reliability of the proposed CNN architecture. Because the dataset was augmented from only one type of Leukemia disease, the acute lymphoblastic leukemia (ALL), we would like to use new dataset including four types of Leukemia to evaluate our architecture. In the future work, we would like to use some algorithm tuning in terms of weight initialization, activation function to improve performance of our CNN architecture.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (No. 2016R1D1A3B03931003, No. 2017R1A2B 2012456) and Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2015-0-00225), and supported by Brain Busan (BB21) project.

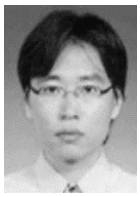
REFERENCE

[1] Melissa Conrad Stöppler, MD. [Online]. Available: https://www.medicinenet.com/leukemia/article.htm#leukemia_facts
[2] C. Nordqvist. [Online]. Available: https://www.medicalnewstoday.com/articles/215538.php?sr
[3] National Cancer Institute. [Online]. Available: https://seer.cancer.gov/statfacts/html/amyl.html
[4] National Cancer Institute. [Online]. Available: https://seer.cancer.gov/statfacts/html/clyl.html
[5] National Cancer Institute. [Online]. Available: https://seer.cancer.gov/statfacts/html/leuks.html
[6] S. Mahazan, S. S. Golait, A. Meshram, and N. Jichlkan, "Review: detection of types of acute leukemia," *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 3, pp. 104-111, 2014.
[7] V. Shankar, M. Deshpande, N. Chaitra, and S. Aditi, "Automatic detection of acute lymphoblastic leukemia using image processing," in *Proc. International Conference on Advances in Computer Applications (ICACA)*, 2016.
[8] I. Vincent, K. R. Kwon, S. H. Lee, and K. S. Moon, "Acute lymphoid leukemia classification using two-step neural network classifier," in *Proc. Workshop on Frontiers of Computer Vision (FCV)*, Mokpo, South Korea, 28-30 Jan. 2015.
[9] T. Ttp, G. N. Pham, J. H. Park, K. S. Moon, S. H. Lee, and K. R. Kwon, "Acute leukemia classification using convolution neural network in clinical decision support system," in *Proc. 6th International Conference on Advanced Information Technologies and Applications (ICAITA 2017)*, Sydney, 2017.
[10] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017.
[11] C. N. Vasconcelos and B. N. Vasconcelos, "Convolutional neural network committees for melanoma classification with classical and expert knowledge based image transforms data augmentation," 2017.
[12] S. Kansal, S. Purwar, and R. K. Tripathi, "Trade-off between mean brightness and contrast in histogram equalization technique for image enhancement," in *Proc. 2017 IEEE International Conference on Signal and Image Processing Applications (IEEE ICSIPA 2017)*, Kuching, Malaysia, 2017.
[13] Padding and shearing an image simultaneously. [Online]. Available: https://www.mathworks.com/help/images/examples/padding-and-shearing-an-image-simultaneously.html

Tran Thi Phuong Thanh received the B.S. degree in information technology from Water Resources University, Vietnam, in 2015. She is currently studying a master course in IT convergence and application engineering at Pukyong National University, Korea. Her research interests include digital signal and image processing, bioinformatics.

Caleb Vununu received the B.S. degree in computer science from Youngsan University, Korea in 2015 and he received an M.S. degree in IT Convergence and application engineering in Pukyong National University, Korea in 2017. He is currently studying Ph.D. course in IT convergence and application engineering in Pukyong National University, Korea. His research interests include digital signal and image processing, bioinformatics.

Sukhrob Atoev received the B.S. degree in telecommunication technologies from Tashkent University of Information Technologies, Uzbekistan, in 2016. He is currently studying M.S. degree in IT convergence and application engineering at Pukyong National University, Korea. His research interests include drone security, drone communication channel, digital signal and image processing.



Suk-Hwan Lee received a B.S., an M.S., and a Ph. D. degrees in electrical engineering from Kyungpook National University, Korea in 1999, 2001, and 2004 respectively. He is currently an associate professor in the Department of Information Security at Tongmyong University. His research interests include multimedia security, digital image processing, and computer graphics.

Ki-Ryong Kwon received the B.S., M.S., and Ph. D. degrees in electronics engineering from Kyungpook National University in 1986, 1990, and 1994 respectively. He worked at Hyundai Motor Company from 1986-1988 and at the Pusan University of Foreign Language from 1996-2006. He is currently a professor in the Dept. of IT Convergence & Application Engineering at the Pukyong National University. He has researched the University of Minnesota in the USA on 2000-2002 with Post-Doc, and Colorado State University on 2011-2012 with visiting professor. He was the general president of Korea Multimedia Society on 2015-2016, also was a director of IEEE R10 Changwon section on 2012-2016. His research interests are in the area of digital image processing, multimedia security and watermarking, bioinformatics, weather radar information processing.