### WHite Blood cells classification

### using Convolutional Neural network

Liu Theodorus David Leonardi (A0178263X), Goh Chung Tat Kenrick (A0080891Y), Rubi Saini (A0178255W), Bhujbal Vaibhav Shivaji (A0178321H), Zheng Kaiyuan (A0054043U), Tan Zhi Wen David (A0056418A)

#### Abstract

Background and objectives: White blood cells are mainly classified into 4 types and their detection, feature extraction and classification is a challenge in the medical field. In the past, cell counting was manual and time-consuming but the introduction of machine learning methods have made the automated counting and classification of white blood cells a potential area where performance can be optimized to greatly reduce the time taken and increase accuracy.

Methods: In this paper, convolutional neural network (CNN) was selected as the machine learning method to extract and classify the four different types of white blood cells. The original images were augmented by different transformations and oversampled to increase their size from 352 to 10,000 so as to provide a well-balanced set of 2,500 images for each of the cell types. We proposed a CNN architecture which was implemented using Keras by Python Programming with TensorFlow at the backend.

Results: Baseline accuracy of 85% was obtained from a paper which used 138 sample images [1]. The CNN model accuracy with and without the image black background was increased from 70.16% and 79.17% to 87.29% and 89.71% respectively using LeNet 5 CNN architecture [2] first and then our proposed model with an additional hidden layer to increase its depth, thus beating the baseline accuracy of 85%. To avoid overfitting, our team decided that one extra layer was enough for our proposed model.

1. **Introduction**

In today's medical landscape, the use of data analytics for medical diagnosis and treatment is becoming commonplace in hospitals, clinics and medical laboratories. Of its many uses, the automated detection, extraction and classification of white blood cell subtypes is a growing field which has the potential to greatly reduce analysis time and manpower while maintaining or improving on the accuracy of identifying and characterizing the white blood cell [3]. While microscopic white blood cell counting by experienced haematologists are accurate as a reference for samples containing abnormal cells, the procedure is time consuming and prone to biasness with low repeatability. Hence, various methods of automating the counting of white blood cells and diagnosing diseases are constantly being studied to better improve the accuracy and duration of counting and recognizing the different cell types [4]. In this paper, the white blood cell is classified into four types, namely the Eosinophil, Lymphocyte, Monocyte and Neutrophil [5]. Eosinophils target multi-cellular parasites and make up around 5% of white blood cells while Lymphocytes target virus infected

and tumour cells and make up around 20% to 45% of white blood cells mainly in the lymphatic system. Monocytes make up about 5% of white blood cells and digest and destroy foreign bodies in a process called "phagocytosis". They are distinguishable by their kidney shaped nucleus and ample cytoplasm. Lastly, the Neutrophil are defenders against bacteria and fungi and are first responders to microbial infection. They make up the largest proportion of white blood cells at 62% and forms pus in huge numbers [6].

In this report, we have selected the machine learning method of Convolutional Neural Network (CNN) to extract the cell features and classify them into the respective four types as mentioned previously [7]. CNN creates learning filters by carrying out multi-convolutions of the same input through multilayers of nodes and is a useful tool for image recognition, segmentation, detection and retrieval. Due to the large amount of time needed to train CNN, a graphic processing unit (GPU) was used as a swift way for parallel computations [8].

1. **LITERATURE SURVEY**

Blood disorders can cause morphological changes in mature red blood cells. In a paper by Mojtaba Taherisadr he and his team investigated blood smears morphologically to study and distinguish blood diseases. This method was based on image processing. It was able to distinguish blood cells into 12 different types [9] [10]. From extracted features one can evaluate if the cells are normal or have some disorders which are helpful in identifying diseases like iron deficiency, anemia, hereditary elliptocytosis or megaloblastic anemia due to folic acid deficiency and other abnormalities [9].

Wu et al. said that in case of blood cell segmentation, edge detection is difficult because all boundaries are not very sharp in images, so it is difficult to locate cell accurately. They developed an iterative, circular histogram-based Otsu’s approach for leukocyte segmentation. This approach was based on least square method [12].

In a paper written by R. Sukesh Kumar, he stated an approach for determination of thresholds over a two-dimensional image histogram using color image segmentation using higher order entropy as a textural feature. Two basic models for color images are the RGB (Red, Green, Blue) color model and the HIS (Hue, intensity, saturation) color model [13].

M. Habibzadeh in his work described a subcomponent system for CBC to automatically detect and classify WBC cells from low resolution cytological images using feature extraction. They used 3 classifiers including SVM and CNN. SVM using features extracted by a kernel principal component analysis of the intensity and histogram features and CNN using entire image as input. All the classifiers were compared, and CNN gave the best results for all five types of WBCs with an accuracy of 87%. [1]

1. **DATA**

**3.1 Data Description**

The white blood cell dataset is obtained from Kaggle dataset repository known as “Blood Cell Images” (<https://www.kaggle.com/paultimothymooney/blood-cells>). Each image contains a type of white blood cells together with red blood cells surrounding it, as shown in Figure 1. Our task is to be able to train a model to make a distinction between the white blood cell from the red blood cell and at the same time be able to identify the type of the white blood cell.

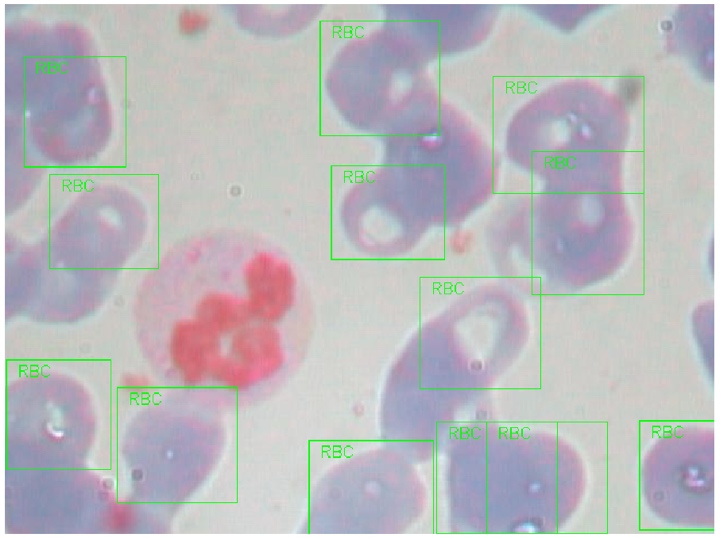


Figure 1. White blood cell surrounded but red blood cell which are tagged by rectangular boundaries.

There is a collection of images of four types of white blood cells, namely the Eosinophil, Lymphocyte, Monocyte, and Neutrophil in the dataset. Sample image of the four types of white blood cell is shown in Figure 2.

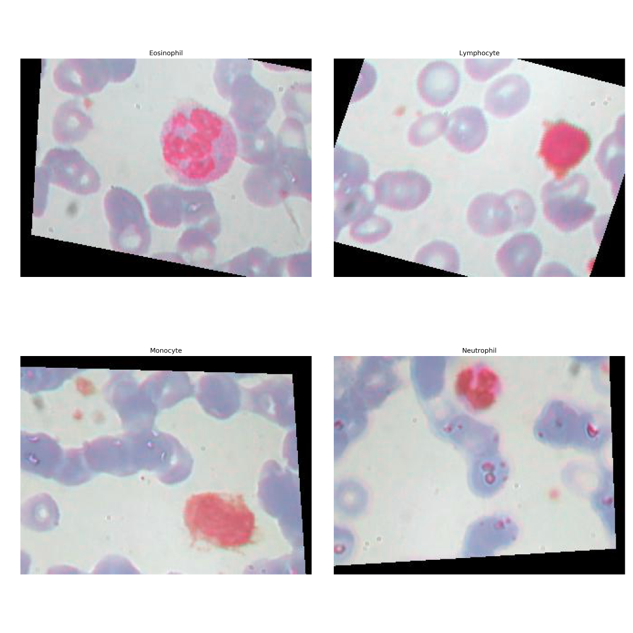


Figure 2. Sample images of four types of white blood cells: Eosinophil, Lymphocyte, Monocyte and Neutrophil.

The images for each type of white blood cells are kept in separate folders and the number of images for each type of white blood cell to serve as inputs to train the model is shown in Figure 3. It can be observed that the dataset is a well-balanced one with equal number of images for each type of white blood cell.

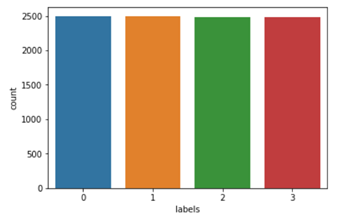


Figure 3. Number of images for each type of white blood cell. Label 1 is Neutrophil, Label 2 is Eosinophil, Label 3 is Monocyte and Label 4 is Lymphocyte.

**3.2 Data Processing**

It is observed that there are borders which are black in color in the images, as a result of image augmentation, which can disrupt the training of the model (i.e. Model could be extracting the border as part of feature extraction). One example of such image with its pixel intensity distribution is plotted in Figure 4. It can be observed that there is a substantial number of pixels with zero intensity.

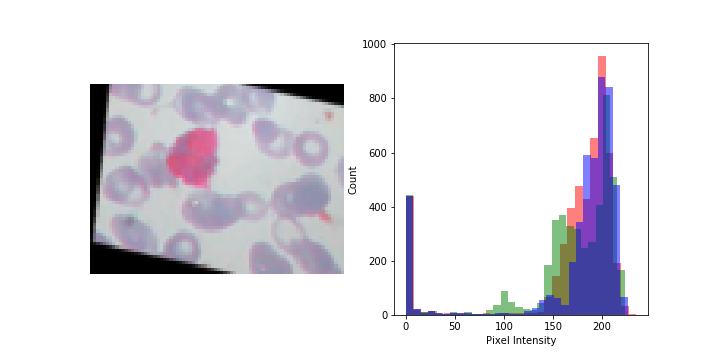


Figure 4. Sample image of white blood cell with its corresponding pixel intensity distribution. It can be seen that there is a substantial number of pixels with zero intensity (black)

In order to work around this problem, we replace the border with the color similar to the background of the main microscopic image, as shown in Figure 5, and the resulting images will then be used for training the model. It can be seen that there are little to no counts of pixels with zero intensity after the replacement.

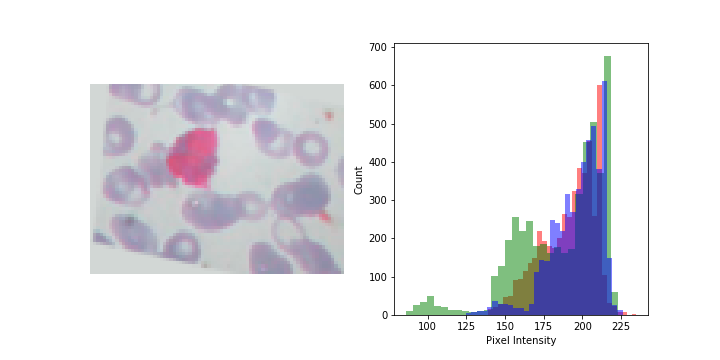


Figure 5. Processed image with the replacement of the border with color similar to the background of the main microscope image

1. **METHOD**

**4.1 Brief Introduction to Convolution Neural Network**

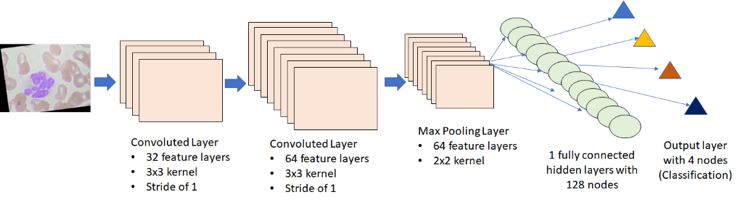
Our approach in classifying the different white blood cells utilizes the Convolution Neural Network (CNN) which was first proposed by Yann LeCun in 1998 [15] and was first successfully applied in the CNN architecture known as the LeNet 5 [2]. The use of CNN in image recognition, which is our core task at hand, is to address two issues. Firstly, images contain a lot of information in the form of pixels. If each individual pixel serves as input into an ordinary fully connected Multiple Layer Perceptron [16], there would be tremendous amount of weights to train that utilized large amount of computational resources. Secondly, since images consist of small number of features such as edges, there would be a need extract them rather than to solely rely on the spatial correlation of pixel values in the image. In the context of our problems, boundary edges of the white blood cells need to be identified so as to distinguish those cells with the background and distinguish among the different type of white blood cells, since different white blood cells would possess different kind of edges. CNN addresses these issues with three main operations: Local Receptive Field, Shared Weights and Sub-Sampling [15], as described briefly in the next paragraph.

The local receptive field (or "Kernel" or "Filter") could be thought of as a matrix of a specific dimension much smaller than the original image within a region in that original image. Only pixels within the field would be connected to a neuron with an output in the layer known as the feature layer. In addition, the weights of all the neurons connecting to the same feature layer would possess the same weight. Essentially, the purpose of this operation is to extract important features (such as edges) from the image and this operation can be repeated to obtain more feature layer with different weights. This set of feature layers is also known as the convoluted layer. Sub-sampling is the operation to encode information into a matrix of smaller dimension derived from the convoluted layer through performing either averaging or obtaining the maximum value within the sample of the convoluted layer. The result is also known as the “Pooling layer”. The purpose of this operation is to reduce the spatial resolution of the convoluted area so as to prevent overfitting. The last few layers in the CNN is typically a fully connected neurons with inputs from the Pooling Layer to output a classification of the image.

We implemented our proposed CNN architecture using Keras [17] via Python programming language, with TensorFlow [18] at the backend. Keras provides the freedom for user to define their own neural network layers. Through this platform, we defined our own CNN architecture specific to the white blood cell identification problem.

**4.2 Baseline CNN Architecture for Identification of White Blood Cell**

Our team has selected this Neural Network as the baseline model as explained in this paper [1]. It is a LeNet-5 architecture and we have used it as our baseline model for comparison [19].  The schematic of this CNN architecture is shown in Figure 6. The model uses the augmented images with black color border as inputs during the model training and testing phase. in this paper [1], we are unable to determine if the author has used pre-trained models or re-training the model from scratch using his training data. Coupled with the fact that our team has more training data than the author, our team has decided to retrain the Lenet-5 model from scratch to use it as the baseline.

Figure 6. Lenet-5 (Baseline) CNN architectural model for comparison

**4.3 Proposed CNN Architecture for Identification of White Blood Cell**

Our proposed CNN architecture is shown in Figure 7. There are three convoluted layers (one after another) responsible in extracting features in the image. The convolution layers contain 32, 64 and 128 filters respectively. This is to ensure that we are able to extract the important features from the images. All use a 3 by 3 kernel with 1 stride. Subsequently, there will be a max pooling layer to reduce the size of the dimension. This is important in our analysis as max pooling remove the spatial relationship within the features itself. Due to this unique feature, our model will be more robust in detecting white blood cells. After which, it will be connected to a fully connected Multiple Layer Perceptron of 1 hidden layers with "ReLu" activation function and 256 nodes, before connecting to the output layer with 4 nodes representing 4 classifications of white blood cells with "SoftMax" activation function. In between, there are "dropout" layers whereby nodes in the layer would be removed randomly to prevent model overfitting. Our team believes that by increasing the number of hidden layers in the architecture, we will be able to extract more precise feature out from the image hereby improving the results from the validation of the model.

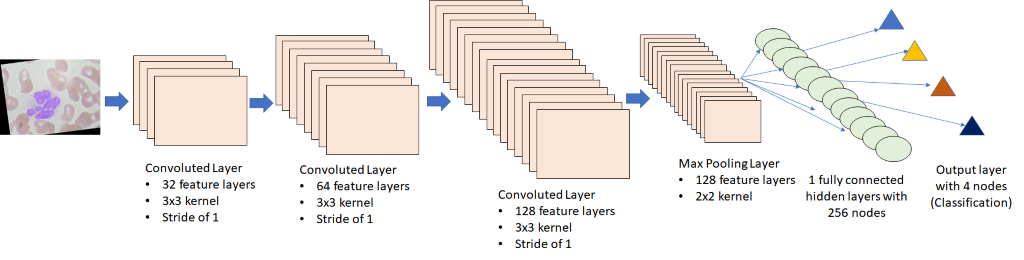


Figure 7. Our proposed CNN architecture used in the identification of the type of white blood cell.

1. **RESULT**

**5.1 Results of the Baseline approach**

Table 1. Performance metric using the baseline approach after 30 epochs.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| **NEUTROPHIL** | 0.52 | 0.72 | 0.60 | 624 |
| **EOSINOPHIL** | 0.51 | 0.36 | 0.42 | 623 |
| **MONOCYTE** | 0.95 | 0.76 | 0.85 | 620 |
| **LYMPHOCYTE** | 0.88 | 0.96 | 0.92 | 620 |
| **Avg/ Total** | 0.71 | 0.70 | 0.70 | 2487 |
| **Test Accuracy** | 70.0165% | | | |

As mentioned in Section 4.2, our team has retrained the LeNet-5 Model from scratch using our training dataset. From Table 1, we can see that there are low precision on both Neutrophil as well as Eosinophil. In addition, there is particularly low recall on Eosinophil. This shows that the model has difficulties in differentiating between Neutrophil and Eosinophil. As such, in our proposed approach, we have increased the number of hidden layers as well as the number of full connected layers at the end to improve the robustness of the model and in turn the accuracy of the validation result.

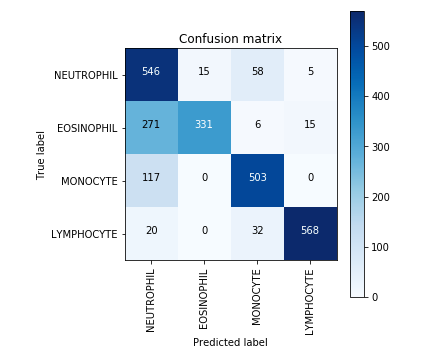


Figure 8. Confusion matrix using the baseline model

**5.2 Proposed approach**

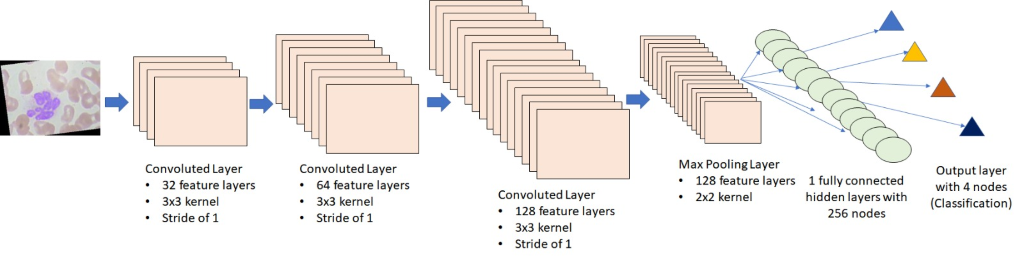


Figure 9. Our proposed CNN architecture

As mentioned in section 5.1, due to the low precision and recall on one of the labels, we propose to increase the depth of our CNN by adding a convolution layer to the base model. Our team has tried different methods in increasing the depth of the CNN. Through our analysis, we have discovered that the result of the model is highly sensitive to the depths of the CNN model. If the CNN model has too many hidden layer, it will result in overfitting and lead to very poor validation results. As such, we have determined the optimal model by increasing a single convolution layer to the base model.

**5.2.1 Feature Engineering of Images**

Through the Data Processing done in section 3.2, we have found significant improvement across the different models that we had run on this dataset. In the table below, we have 4 models, namely the base model, our model as well as the pretrained models VGG and LeNet. This table showed us that the quality of the image does play a part in improving the results during the training of the model. We hypothesize that the CNN model is able to extract the unique features of the blood cells more effectively due to the non-presences of the "outlier" black patches at the side of the images, hereby leading to better results. As such, for the subsequent modelling, we will be using the edited photos for the tuning of the hyper parameters of the different model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Accuracy** | **Precision** | **Recall** | |
|  |  |  |

Figure 10. Performance Metric using images with black and without black border

* + 1. **Tuning of Hyper Parameters**

In order to optimize our model further, we ran grid search with these parameters:

* Optimizer: Adadelta
* Epochs: 30, 50
* Learning rate: 1.0, 0.8
* Learning Rate Decay: 0.0, 0.001

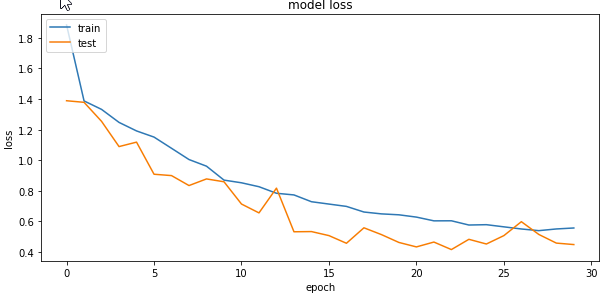
Through our tuning of our Hyper Parameters, we have found out that the results converges rapidly in 30 epochs. Increasing the number of epochs will result in in overfitting of the data. In addition, we have tweaked with the learning rate as well as decay to determine the optimal value for the best accuracy. We have also program the code to choose the weights which provides the best test accuracy instead of the weights of the last epochs. This is to ensure that we have the optimal model with the best test results.

Figure 9. Loss function graph of our tuned model

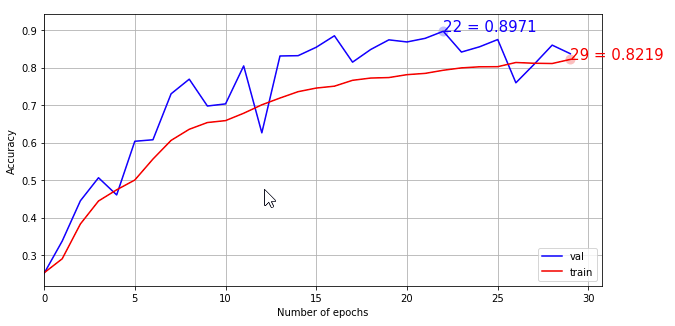


Figure 10. Training and Test accuracy of our tuned model

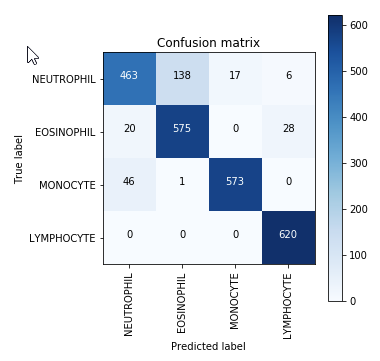


Figure 11. Confusion Matrix of our tuned Model

From these figures, we can see that our tuned model has low loss test function with high accuracy results. Coupled with low number of epochs, we have a robust model that is able to predict the different type of blood cells properly

**Discussion and Conclusion**

**6.1 Conclusion**

We classified the 4 types of white blood cells first using the LeNet 5 architecture as our baseline model and then further improve on the architecture by adding an extra convolution layer. The performance of our propose model improved from the baseline architecture tremendously.

**6.2 Limitation**

The main limitation of a CNNs is that, they can predict good only train and test images have same alignment. CNN is bad at encoding different representations of pose and orientation within themselves. To eliminate this limitation, we have to provide augmented set of images of training image while training which increases the computational complexity and time. A CNN makes predictions by looking at an image and then checking to see if certain components are present in that image or not. If they

are, then it classifies that image accordingly. So even if that component belongs to some other image, CNN will classify it as image from which this feature has been extracted. In a CNN, all low-level details are sent to all the higher-level neurons. These neurons then perform further convolutions to check whether certain features are present. This is done by striding the receptive field and then replicating the knowledge across all the different neurons. Instead of having the information go through all the neurons, like in a conventional CNN, it is better to route the image to specific neurons that have the capability to deal with those features. This will help to reduce the overall time complexity and time required for 1 epoch.

**6.3 Further works**

The novel approach with proposed CNN model can be used to classify several types of Haemophilia which are [haemophilia A](https://en.wikipedia.org/wiki/Haemophilia_A), [haemophilia B](https://en.wikipedia.org/wiki/Haemophilia_B" \t "_blank), [haemophilia C](https://en.wikipedia.org/wiki/Haemophilia_C" \t "_blank), parahaemophilia, and acquired haemophilia A. [20]. These different types of Haemophilia are caused due to lack of different blood clotting factors. Our model can be used to identify those clotting factors and subsequently classify the type of Haemophilia.

1. **REFERENCE**

|  |  |
| --- | --- |
| [1] | M. Habibzadeh, A. Krzyżak and T. Fevens, "White Blood Cell Differential Counts Using Convolutional Neural Networks for Low Resolution Images," *International Conference on Artificial Intelligence and Soft Computing,* pp. 263-274, 2013. |
| [2] | Y. LeCun, "LeNet-5, convolutional neural networks," 2015. [Online]. Available: http://yann.lecun.com/exdb/lenet. |
| [3] | M. C. Su, C. Y. Cheng and P. C. Wang, "A neural-network-based approach to white blood cell classification.," *The scientific world journal,* 2014. |
| [4] | J. Zhao, M. Zhang, Z. Zhou, J. Chu and F. Cao, "Automatic detection and classification of leukocytes using convolutional neural networks," *Medical & biological engineering & computing,* vol. 55, no. 8, pp. 1287-1301, 2017. |
| [5] | P. S. Hiremath, P. Bannigidad and S. Geeta, "Automated identification and classification of white blood cells (leukocytes) in digital microscopic images," *IJCA special issue on “recent trends in image processing and pattern recognition” RTIPPR,* 2010. |
| [6] | M. S. Blumenreich and C. Methods:, "The white blood cell and differential count," *The History, Physical, and Laboratory Examinations [Internet]. 3rd ed. Boston: Butterworths.,* 1990. |
| [7] | A. I. Shahin, Y. Guo, K. M. Amin and A. A. Sharawi, "White blood cells identification system based on convolutional deep neural learning networks," *Computer methods and programs in biomedicine,* 2017. |
| [8] | "A Beginner's Guide to Deep Convolutional Neural Networks (CNNs)," [Online]. Available: https://deeplearning4j.org/convolutionalnetwork. |
| [9] | M. Taherisadr, M. Nasirzonouzi, B. Baradaran, A. Mehdizade and I. Shiraz, "New approch to red blood cell classification using morphological image processing," *Shiraz E-Medical Journal,* pp. 44-53, 2013. |
| [10] | R. Adollah, M. Y. Mashor, N. F. Mohd Nasir, H. Rosline, H. Mahsin and H. Adilah, "Blood Cell Image Segmentation: A Review," *Abu Osman N.A., Ibrahim F., Wan Abas W.A.B., Abdul Rahman H.S., Ting HN. (eds) 4th Kuala Lumpur International Conference on Biomedical Engineering 2008. IFMBE Proceedings, vol 21. Springer, Berlin, Heidelberg,* 2008. |
| [11] | P. Bamford, "Empirical comparison of cell segmentation algorithms using an annotated dataset.," *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on (Vol. 2, pp. II-1073). IEEE.,* 2003. |
| [12] | J. Wu, P. Zeng, Y. Zhou and C. Olivier, "A novel color image segmentation method and its application to white blood cell image analysis," *Signal Processing, 2006 8th International Conference on (Vol. 2). IEEE.,* 2006. |
| [13] | R. S. Kumar, A. Verma and J. Singh, "Color image segmentation and multi-level thresholding by maximization of conditional entrophy," *International Journal of Signal Processing,* vol. 3, no. 1, pp. 121-125, 2006. |
| [14] | K. Jiang, Q. M. Liao and S. Y. Dai, "A novel white blood cell segmentation scheme using scale-space filtering and watershed clustering," *Machine Learning and Cybernetics, 2003 International Conference on (Vol. 5, pp. 2820-2825). IEEE.,* 2003. |
| [15] | Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE,* vol. 86, no. 11, pp. 2278-2324, 1998. |
| [16] | F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychological review,* vol. 65, no. 6, p. 386, 1958. |
| [17] | F. Chollet, "Keras," 2015. [Online]. Available: https://keras.io. |
| [18] | M. Abadi and et.al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems.," *arXiv preprint arXiv:1603.04467,* 2016. |
| [19] | P. T. Mooney, "Identify Blood Cell Subtypes from Images.," Kaggle, 2018. [Online]. Available: https://www.kaggle.com/paultimothymooney/identify-blood-cell-subtypes-from-images. |
| [20] | V. S. Blanchette, N. S. Key, L. R. Ljung, M. J. Manco‐Johnson, H. M. Berg and A. Srivastava, "Definitions in hemophilia: communication from the SSC of the ISTH," *Journal of Thrombosis and Haemostasis,* vol. 12, no. 11, pp. 1935-1939, 2014. |
| [21] | K. B. How, A. S. K. Bin, N. T. Siong and K. K. Soo, "Red blood cell segmentation utilizing various image segmentation techniques," *Proceeding of International Conference on Man-Machine Systems,* pp. 15-16, 2006. |
| [22] | M. Wang, X. Zhou, F. Li, J. Huckins, R. W. King and S. T. Wong, "Novel cell segmentation and online learning algorithms for cell phase identification in automated time-lapse microscopy," *Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on (pp. 65-68). IEEE.,* 2007. |