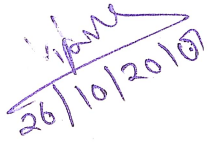


Early Cancer Detection

Arun K, Kamal K, Sivajee Raju N, Sanyam H, Samhith

Introduction

Leukemia, a type of cancer found in your blood and bone marrow, is caused by the rapid production of abnormal white blood cells caused by radiation exposure [Stoppler, 2017]. The high number of abnormal white blood cells are not able to fight infection, and they impair the ability of the bone marrow to produce red blood cells and platelets. 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 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 occurs in a type of marrow cell that forms lymphocytes. Myelogenous leukemia 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) [Nordqvist, 2017].

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. 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. [National Cancer Institute., 2015a] . 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%. [National Cancer Institute., 2014]. Chronic myeloid leukemia occurs mainly in adult with the five-year survival rate is 66.9%

About 5,960 new cases of ALL (3,290 in males and 2,670 in females) every year resulting in 1,470 deaths (830 in males and 640 in females) [American Cancer Society, 2014]. 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. [National Cancer Institute., 2015b] Clinical diagnosis of Acute leukemia is based on a bone marrow examination by a pathologist, the test result is based on the experience of the technician which involves a human judgement and error factor and is time taking. Therefore, using an automatic system to early diagnosis leukemia eliminates this factor has an important role in Leukemia diagnosis. Now, in the world of such advance technology which proves itself as our friend, we can start the identification of blood disorders through visual inspection of microscopic images of blood cells.

From the identification of blood disorders, it can lead to classification of certain diseases related to blood. This paper describes a preliminary study of developing a detection of leukemia types using microscopic blood sample images. [T. T. P. Thanh and Kwon, 2018] Blood is the main source of information that gives an indication of changes in health and development of specific diseases. Changes in the number or appearance of elements that formed will guide health condition of an individual. Analysing through images is very important as from images, diseases can be detected and diagnosed at earlier stage. From there, further actions like controlling, monitoring and prevention of diseases can be done. Images are used as they are cheap and do not require expensive testing and lab equipment.

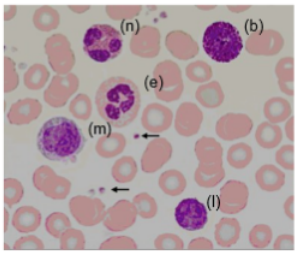


Fig 1 Blood white cells marked with colorant: basophil (b), eosinophil (e), lymphocyte (l), monocyte (m), and neutrophil (n). Arrows indicate platelets. Others elements are red cells. [T. T. P. Thanh and Kwon, 2018]

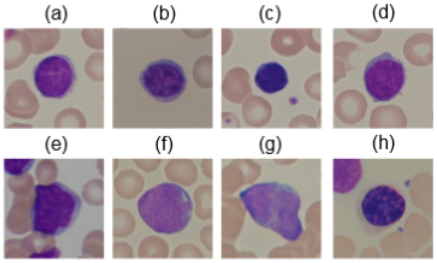


Fig. 2 Example images contained in the ALL-IDB2: healthy cells from non-ALL patients (a-d), probable lymphoblasts from ALL patients (e-h). [T. T. P. Thanh and Kwon, 2018]

The system focuses on white blood cells. The system uses features in microscopic images of peripheral blood sample smear slides and examine changes on texture, geometry, colour and statistical analysis to differentiate a healthy white blood cell (lymphocyte) from a cancerous cell [Ruggero Donida Labati, 2010]. Acute lymphocytic leukaemia can be detected by microscopic inspection of peripheral blood samples. The inspection consists in the search of white cells with malformations due to the presence of a cancer [Walter, 2012]. For decades, this task has been performed by experienced operators, which basically perform two main analyses: Cell classification and Counting.

Related Works

Researches about leukemia classification in recent years are mainly based on computer vision techniques [S. Mahazan, 2014], [V. Shankar and Aditi, 2016]. 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 [I. Vincent and Moon, 2015]. 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.

W.Shitong and Min [January 2006] developed a technique merging the threshold segmentation, fuzzy and some mathematical morphology. It is very good in detecting the leukocytes faster than any other technique does. The problem in this technique is that it is not separating the nucleus and cytoplasm properly. M.Ghosh and Ray [2010] introduced the technique to find out the accurate threshold for segmentation of the leukocytes. He used fuzzy divergence in that technique. He used various functions like Gaussian, Gamma, Gauchy etc. in that technique. This technique works well for segmenting the nucleus but the extraction of cytoplasm has not been taken care of which is also as important as nucleus extraction in cancer detection. L.B. Dorini and Leite. [2007] proposed a scheme for nucleus extraction. The watershed transform has been used in this scheme which is based on the image forest transform. He has extracted the cytoplasm by using the size distribution information. This scheme does not work well if the cytoplasm isnt round.

Angulo and G.Frandlin [2003] proposed a system in which he proposed two-stage blood image segmentation algorithm. They are using binary-filtering and some automatic threshold techniques. This system performs well for extracting the nucleus, cytoplasm and the nucleolus from the lymphocyte images. The two stage segmentation process has been applied here and because of this computation time is higher. The images are taken under different lightening condition which makes difficult to choose the optimum threshold for segmentation. H.J Escalante and Rosales [2012] invented a scheme for classifying leukemia using the swarm model. The leukemia cells need to be isolated manually to make this system work. These isolated cells are then segmented by Markov random fields. These segmented nucleus and cytoplasm are then used to find out features of the type of leukemia

In the work proposed by M. Madhukar and Chronopoulos [1953], a system was developed to detect ALL using images from a single database. These images have multiple nuclei per image. The pre-processing step consists in the conversion of the image from RGB to the L*a*b colour space. In the segmentation step, the unsupervised algorithm K-means is applied to components *a and *b from the converted image, with the number of groups equals to three. In the feature extraction step, they used shape features (e.g. area, perimeter, compactness, solidity, eccentricity, elongation and form-factor), Grey Level Co-occurrence Matrix (GLCM) [R. Haralick and Dinstein, 1973] and Fractal dimension [Poggio and Krotkov, 1992] as descriptors. In order to evaluate the system, they used 98 blood images from the ALL-IDB1 database [R. D. Labati and Scotti, 2011]. The classification was performed using Support Vector Machine and three techniques for

cross-validation: k-fold, Hold-Out and Leave-One-Out. After analysing the results, the authors concluded that the technique that obtained the best accuracy 93.50%.

I. Vincent and Moon [2015] proposed the use of neural networks as classifiers. The method proposed in this work starts converting the image from RGB to the L*a*b colour space. The resulted image is used in the clustering algorithm k-means, which separates the image into three different classes based on their colour information. Contrast enhancement, auto-thresholding and morphological operations are applied in order to obtain the nucleus segmented image. feature extraction and classification stages are subdivided into two steps. The first feature vector set obtained consists of five textural features, four Grey Level Co-occurrence Matrix (GLCM) features (e.g. energy, entropy, contrast, and correlation) and one fractal feature which is represented by Hausdorff Dimension. These feature vectors are analysed by PCA algorithm which produces the input source for the first neural network classifier, the purpose of this classifier is to classify the cells in normal and abnormal. The same algorithm is applied to the second feature extraction process, though the extracted features are different.

Since the second classifier needs a better differentiation, five geometrical features (e.g. cell area, nucleus area, cytoplasm area, nucleus-to-cytoplasm area ratio, and nucleus-to-cell area ratio) are extracted and analysed by PCA algorithm in order to produce input for second neural network classifier that identifies AML and ALL. Both neural networks are trained using Levenberg-Marquardt (LM) algorithm [Lourakis, 2005]. The system achieved an accuracy of 97.70% using 100 blood images from ALL-IDB1 base.

Patel and Mishra [2015], the authors presented an automatic system for the detection of leukemia using microscopic blood images. This work can be divided into pre-processing, segmentation, feature extraction and classification stages. In the first stage, filters were used to remove possible noises in the image, to facilitate the segmentation of the image. The authors, unlike other state of-the-art works, do not make changes in the colour space, using the original RGB colour space. In the segmentation step, the image is converted to grayscale and the clustering algorithms K-means and G. W. Zack and Latt [1977] are applied. In the feature extraction stage, colour, geometry, texture and statistics features were used. ALL-IDB1 is used to evaluate this system, however, only 27 images were used in the tests. The system achieved an accuracy of 93.57% using SVM.

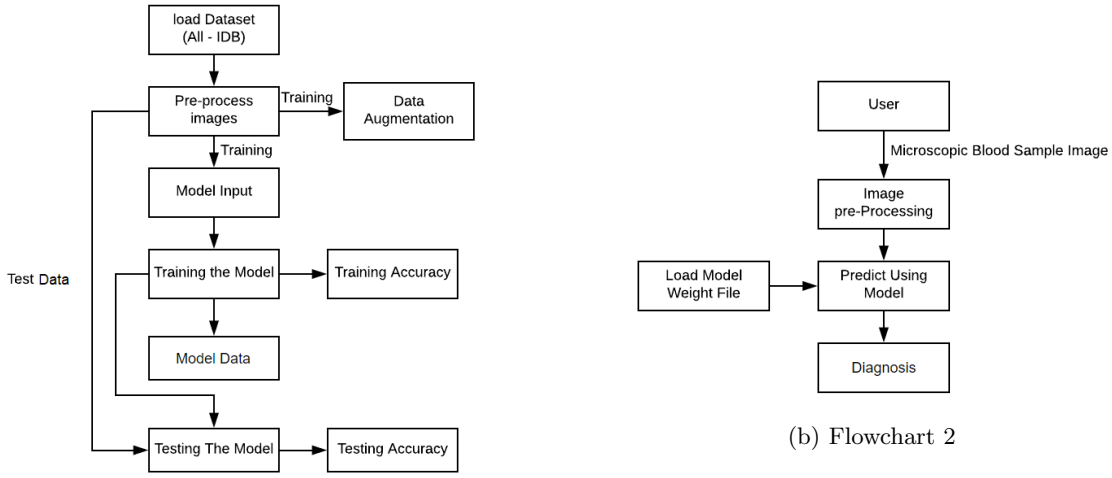
The system proposed by S. Agaian, M. Madhukar, A. T. Chronopoulos, 2012 presented an approach for the classification of blood images with multiple nuclei. The authors converted the images from RGB to the L*a*b color space and applied the clustering algorithm Kmeans. The chosen features were: shape, color, GLCM, Haar wavelet and Fractal dimension. SVM was used as the classifier. The system obtained an accuracy above 94.00% using 98 images from ALL-IDB1.

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 classifications 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 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.

Proposed Methodology

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 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 image 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.



(a) Flowchart 1

(b) Flowchart 2

Fig. 3 Flowcharts

Architecture

As shown in Fig. 3, in this work, we use a network containing 9 layers, as we see in Figure. The first 6 layers perform feature extraction and the other 3 layers (2 fully connected and 1 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 filter size 3x3 and number of filters, 32 and 64, respectively. In the convolution layer 4, we used a filter size of 3x3 and a total of 128 filters. The stride is 2 and no zero-padding was applied.We used max pooling layers with filter size 2, stride 2 to decrease the volume spatially between

convolutional layers 1, 2 and layers 2, 3. During the learning, the chosen size of the mini batch was 100. ReLu is used as the activation function. We used two fully connected layers of sizes 128 and 2.

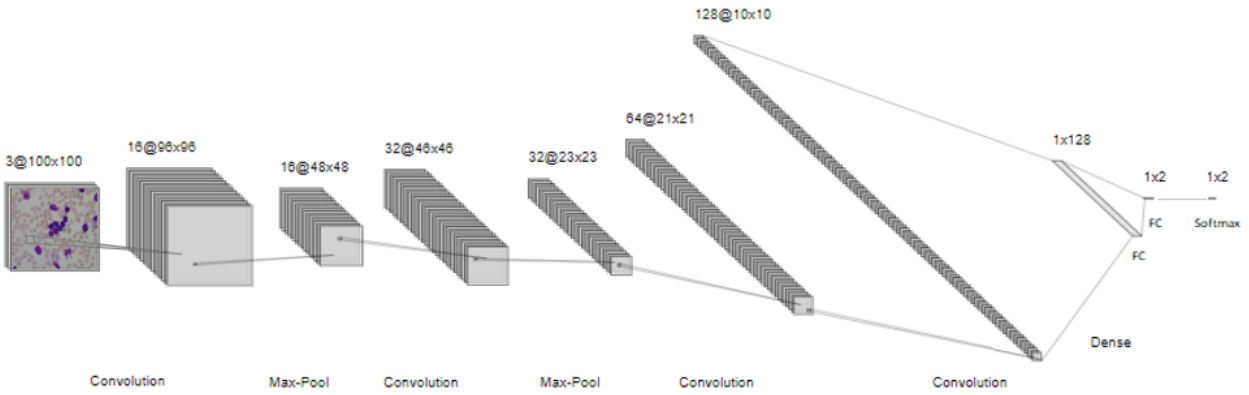


Fig. 4 Architecture of the model

In the preliminary version of this work that has been published in the 6th International Conference on Advanced Information Technologies and Applications (ICAITA 2017) [Thanh T.T.P. and K.R.Kwon, 2017], 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. We present a deeper architecture trained on a significantly augmented dataset.

Algorithm

Import packages: keras, numpy, os, PIL, theano, scikit-learn, scikit-image
Change the directory of the program to location of dataset: ALL-IDB
imgfiles = All images in the directory
declare x,y as empty lists
for all images in imgfiles **do**
    im= (load image) in RGB format
    im= resize image to 100x100
    imrs= convert im to numpy array; transpose and reshape
    perform data augmentation techniques and append the dataset
    append image into x and image class(from file name) into y
end for
Split the dataset in 4:1 for training and testing
Create a sequencial model
Add model layers (Refer Architecture)
Fit and train the model using training dataset(prints epochwise training accuracy)

16: Test the model(prints testing accuracy)

17: Save the model weights into h5 file

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, we used some simple data augmentation methods to expand the available data. [Perez and Wang, 2017], [Vasconcelos and Vasconcelos, 2017].

1. Histogram Equalization

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[S. Kansal and Tripathi, 2017].The equalized image shows more detail of white blood cells and red blood cells in the ALL image.

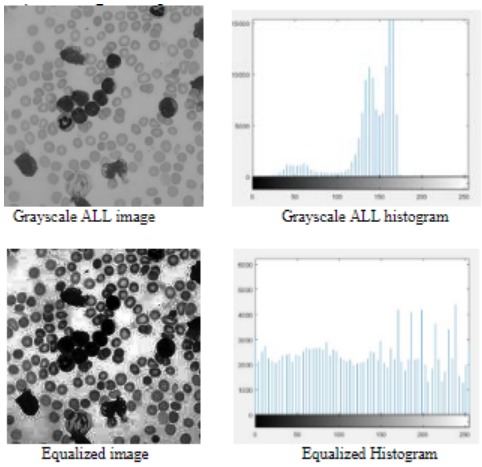


Fig. 5 The grayscale ALL image and the adjusted image [T. T. P. Thanh and Kwon, 2018]

2. Translation

In this paper, we 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.

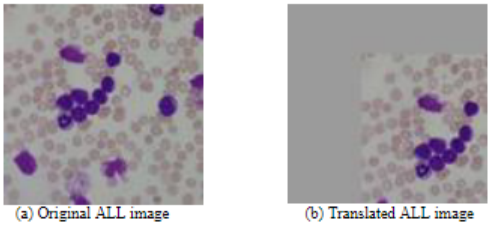


Fig. 6 The original image and the translated image [T. T. P. Thanh and Kwon, 2018]

3. Reflection

Image reflection is a method to mirror an image through the vertical and horizontal axis. This is a simple solution to extend dataset but improves to high efficiency because the mirrored image still has the content of the image.

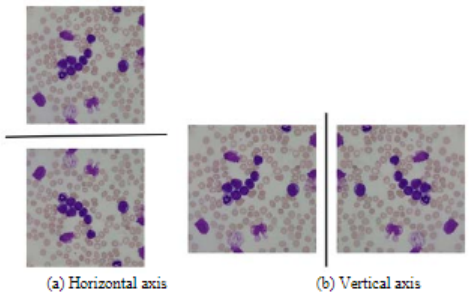


Fig. 7 Image reflection through vertical and horizontal axis [T. T. P. Thanh and Kwon, 2018]

4. 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][MathWorks Documentation, 2017].

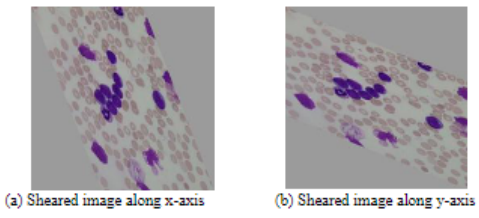


Fig. 8 The original ALL image and sheared image [T. T. P. Thanh and Kwon, 2018]

5. Grayscale Image

We convert all of the images in ALL-IDB1 dataset from RGB format to grayscale image.

6. Blurring Image (RGB and Grayscale)

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

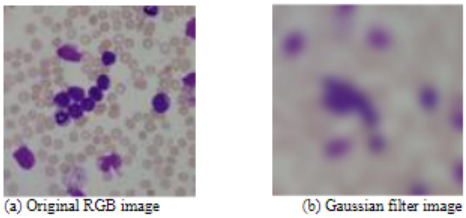


Fig. 9 The original ALL image and Gaussian filtered image [T. T. P. Thanh and Kwon, 2018]

7. Rotating Image

Original image is rotated a rotation angle between[180, 180] 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 are set to 155. After rotating the image, resize the output image to be the same size with the input image.

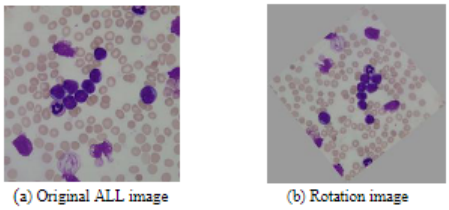


Fig. 10 The original image and rotated image [T. T. P. Thanh and Kwon, 2018]

Experimental Results

The ALL-IDB1 database consists of 108 microscopic peripheral blood sample images of which 59 have normal lymphocytes and 49 have blast cells. In this paper, to increase the accuracy of our model, we have augmented the dataset to 1080 images by applying image transformations mentioned earlier in this paper. Our research was conducted on python using keras. 70% of the images (756) were used for training and the rest were used for testing and validation. For the verification of the results, we chose k-fold cross-validation, with the value of k being 5. This value was chosen because it presented a larger set of images to be validated. Validation accuracy obtained was 97.6% and loss was 0.066. We have presented a deeper CNN model as compared to the [Thanh T.T.P. and K.R.Kwon, 2017] and changed the size of the input in order to improve the recognition of leukemia (our model has achieved a testing accuracy of 95.2%) As shown in the table, our proposed CNN architecture of the 189 actual normal cell images, it predicted that 185 images were normal cell images and 4 were blast cell images, and of the 168 abnormal cell images, it predicted that 156 images were blast cell images and 12 as normal cell images, as shown below in the confusion matrix.

	Predicted Normal	Predicted Abnormal
Actual Normal	184	5
Actual Abnormal	12	156

Confusion Matrix

	Precision	Recall	f1-score	Support
Normal Cell	0.94	0.97	0.96	189
Abnormal Cell	0.97	0.93	0.95	168
avg/total	0.95	0.95	0.95	357

Classification Report

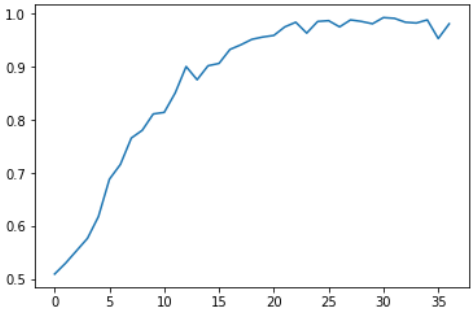


Fig. 11 Training Accuracy vs Epochs

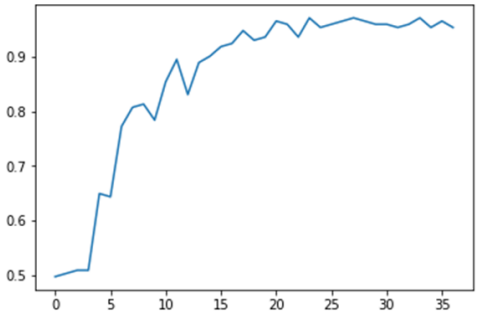


Fig. 12 Validation Accuracy vs Epochs

Comparison with other works

S.no	Author year	Method	Classification accuracy (%)
1	Himali P. Vaghela, Hardik Modi, Manoj Pandya, M.B. Potdar	Edge detection using histogram equalizing method and linear contrast stretching	73.70%
2	Karthikeyan and Poornima, 2017	Image segmentation using fuzzy C-means, K means, SVM	90.00%
3	M. Madhukar, S. Agaian, and A. T. Chronopoulos, 2012	Unsupervised algorithm K-means, SVM, cross validation techniques	93.50%
4	N. Patel and A. Mishra, 2015	K-means and Zack, SVM.	93.57%
5	Mohammad Ehtasham Billah,2018	BCNN, mc dropout	94.00%
6	Mohapatra et al., 2014	Ensemble classifier (Naïve Bayesian, K-nearest neighbour, multilayer perceptron)	94.73%
7	T.T.P. Thanh, Caleb Vununu, Sukhrob Atoev, Suk Hwan Lee, and Ki Ryong Kwon	CNN	96.60%
8	I. Vincent, K.-R. Kwon, S.-H. Lee, and K.-S. Moon, 2015	Neural networks as classifiers	97.70%
9	**Proposed**	**CNN, K-fold**	**99.12%**

Comparison with related works

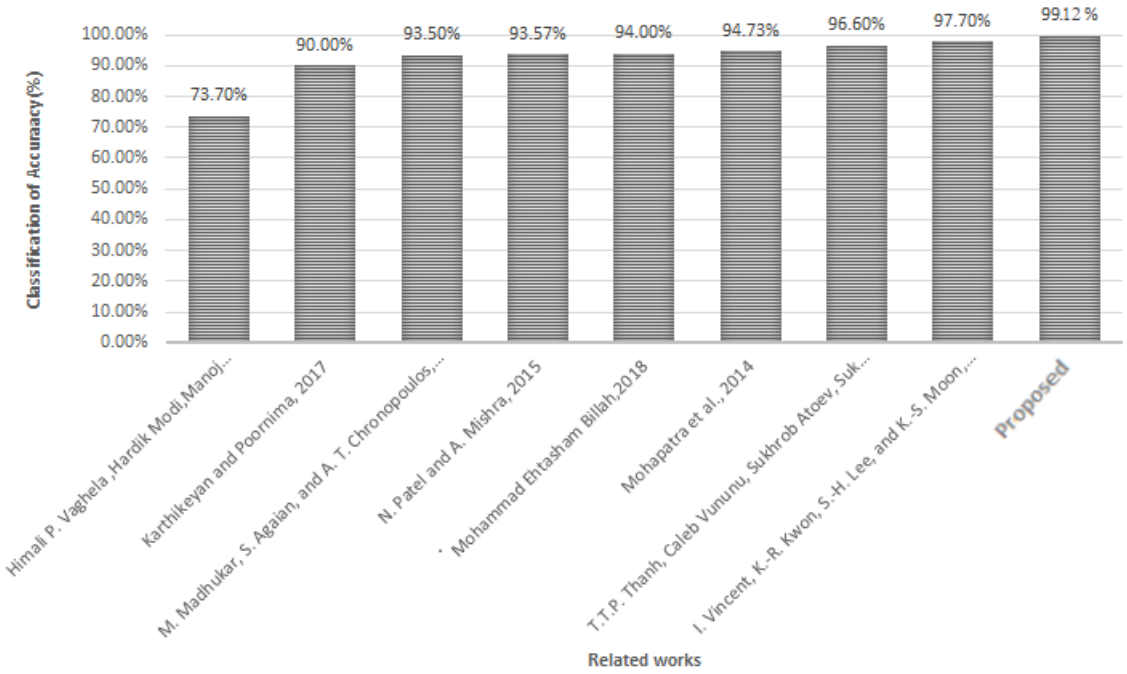


Fig. 13 Comparison of training accuracies

Conclusion and future works

Our aim was to develop a different version of CNN which can be used to detect leukemia from ALL images. We not only focused on getting predictions from the models we have created but also developed the model to be more reliable in producing accurate predictions This method of white blood cell classification is promising and can be used in diagnostic systems for leukemia for early detection of the disease. The work presented in this paper describes a new system for the diagnosis of leukemia in blood sample images using a Convolutional Neural Network and cross validation technique namely k fold. The dataset we worked on is relatively small and can lead to overfitting. Hence we have performed the proposed method on a dataset which is largely augmented in order to confirm the accuracy and reliability of the proposed CNN architecture. Based on the results obtained, it is possible to say that this methodology has been robust in comparison with others.

In the future work, we mean to deliver an automated system which can advise prognosis and course of treatment, which not only can classify if the blood sample is cancerous, but also classify leukocytes and count the number of blast cells in the sample using the YOLO model. We propose the use of fine tuning in the architecture in terms of weight initialization and activation function.

In addition, we intend to use new image databases of the three other types of leukemia from research driven cancer treatment facilities, so the system can be used in daily life, helping doctors and patients in the wholesome diagnosis of this disease.

References

Melissa Conrad Stoppler. Leukemia. 2017.

C. Nordqvist. What you need to know. *Acute Myeloid Leukemia*, 2017.

National Cancer Institute. Cancer stat facts. *Leukemia- Acute Myeloid Leukemia (AML)*, 2015a.

National Cancer Institute. Cancer stat facts. *Leukemia- Chronic Lymphocytic Leukemia(CLL)*, 2014.

American Cancer Society. Key statistics for acute lymphocytic leukemia. *Lymphocytic Leukemia*, 2014.

National Cancer Institute. Cancer stat facts. *Leukemia*, 2015b.

Sukhrob Atoev Suk-Hwan Lee T. T. P. Thanh, Caleb Vununu and Ki-Ryong Kwon. Leukemia blood cell image classification using convolutional neural network. *International Journal of Computer Theory and Engineering*, 10(2), 2018.

Fabio Scotti Ruggero Donida Labati, Vincenzo Piuri. The acute lymphoblastic leukemia image database for image processing. 2010.

John Walter. Understanding leukemia. *Leukemia and Lymphoma society*, 2012.

A. Meshram andN. Jichlkan S. Mahazan, S. S. Golait. detection of types of acute leukemia. *International Journal of Computer Science and Mobile Computing*, 3(3):104–111, 2014.

N. Chaitra V. Shankar, M. Deshpande and S. Aditi. Automatic detection of acute lymphoblastic leukemia using image processing. *International Conference on Advances in Computer Applications*, 2016.

SH. Lee I. Vincent, KR. Kwon and KS. Moon. Workshop on frontiers of computer vision. *Acute Lymphoid Leukemia Classification using Two Step Neural Network Classifier*, 2015.

W.Shitong and W. Min. Ieee transactions on information technology in biomedicine. *A new algorithm based on fuzzy cellular neural networks for white blood cell detection.*, January 2006. doi: 10.1016/S0031-8914(53) 80099-6.

C.Chakraborty M.Ghosh, D.Das and A.K Ray. Automated leukocyte recognition using fuzzy divergence. 2010.

R.Minetto L.B. Dorini and N.J Leite. Proceeding of the brazilian symposium on computer graphics and image processing. *White blood cell segmentation using morphological operators and scale-space analysis*, pages 294–304, 2007.

J. Angulo and G.Frandlin. Microscopic image analysis using mathematical morphology: Application to haematological cytology. *Mendez-Vilas, editor ,Science ,Technology and Education of Microscopy*, pages 304–312, 2003.

J.A Gonzalez P.G.Gil L.Altamirano C.A.Reyes-C.Reta H.J Escalante, M.M Gomez and A. Rosales. Acute leukemia classification by ensemble particale swarm model selection. *Artificial Intelligence In Medicine*, 2012.

S. Agaian M. Madhukar and A. T. Chronopoulos. New decision support tool for acute lymphoblastic leukemia classification. 1953. doi: 82951882951812.

K. Shanmugam R. Haralick and I. Dinstein. Texture features for image classification,. *IEEE Transactions on Systems, Man, and Cybernetics,*, 3(6), 1973.

K. Arakawa Poggio, Koch and E. Krotkov. A. pentland. fractal-based description of natural scenes. 1992.

V. Piuri R. D. Labati and F. Scotti. All-idb: The acute lymphoblastic leukemia image database for image processing. *Image Processing (ICIP), 2011 18th IEEE International Conference*, 2011.

M. I. A. Lourakis. A brief description of the levenberg-marquardt algo- rithm implemented by levmar. 2005. doi: 10.1016/S0031-8914(53)80099-6.

N. Patel and A. Mishra. Automated leukaemia detection using microscopic images,. 58, 2015.

W. E. Rogers G. W. Zack and S. A. Latt. Automatic measurement of sister chromatid exchange frequency,. *Journal of Histochemistry & Cytochemistry,*, page 741753, 1977.

J.H. Park K.S.Moon S.H.Lee Thanh T.T.P., G.N.Pham and K.R.Kwon. Acute leukemia classification using convolution neural network in clinical decision support system. *Proc. 6th International Conference on Advanced Information Technologies and Applications (ICAITA 2017),Sydney*, 2017.

L. Perez and J. Wang. The effectiveness of data augmentation in image classification using deep learning,. 2017.

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.

S. Purwar S. Kansal and R. K. Tripathi. Trade-off between mean brightness and contrast in histogram equalization technique for image enhancement,. *2017 IEEE International Conference on Signal and Image Processing Applications (IEEE ICSIPA 2017)*, 2017.

MathWorks Documentation. Padding and shearing an image simultaneously. 2017.