1. INTRODUCTION

Blood consists of plasma, and three different types of cells and they are: White Blood Cells, Red Blood Cells and Platelets and each of these performs particular task. Red blood cells transport oxygen from the lungs to the tissues of the body and vice versa. White blood cells help the body to fight against diseases and infections. Platelets help to clot and control bleeding. Leukemia is cancer of blood cells in which number of white cells is increases numerously and those are immature cells that interfere with other blood cells, usually red blood cells and platelets. Our body's white blood cell ratio is 1000:1. It means that between 1000 red blood cells there is 1 white blood cell.

There are two types of white blood cells that get turn into leukemia and they are:

1. Lymphoid cells
2. Myeloid cells

Leukemia that caused due to lymphoid cells is called lymphocytic or lymphoblastic leukemia and if it is caused due to myeloid cells then it is known as myelogenous or myeloid leukemia. Leukemia is grouped in two ways: acute or chronic, grouped according to how fast the cells are growing. The abnormal blood cells in acute leukemia are usually immature blasts (young cells) that do not work properly. These cells are growing fast. Acute leukemia gets worse quickly unless it is immediately treated. Young blood cells are present in chronic leukemia, but also mature functional cells are produced. Blasts are growing slowly in chronic leukemia. It takes the disease longer to get worse.

The four major forms of leukemia are:

1. Acute lymphoblastic leukemia (ALL)
2. Acute myelogenous leukemia (AML)
3. Chronic lymphocytic leukemia (CLL) and
4. Chronic myelogenous leukemia (CML)

1.1 PROBLEM DEFINITIONS

According to the Leukemia & Lymphoma Society, one person in the U.S. is diagnosed with blood cancer approximately every 3 minutes and an estimated total of 174,250 people in the U.S. are expected to be diagnosed with leukemia, lymphoma or myeloma in 2018. The estimated new cases in 2019 are around 61,780 and according to the National Cancer Institute, the percentage of all new cancer cases is 3.5 percent. As in

acute leukemia, if the treatment is not done in a precise time, the person died within a few months. And it is very necessary to detect cancer in the early stages to treat this type of cancer or any type of cancer. It takes more time and effort to do the detection process by technicians manually and it costs more with the help of the instrument.

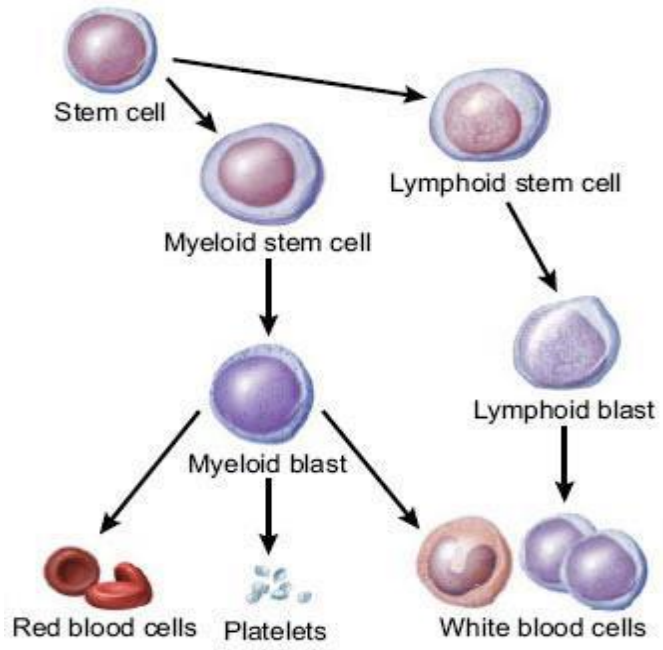


Figure1(a): The formation of Myeloid and Lymphoid series of cell.(researchgate.net)

1.2 PURPOSE

The purpose of our project is to develop a system that can automatically detect cancer from the blood cell images. This system uses a convolution network that inputs a blood cell images and outputs whether the cell is infected with cancer or not. The appearance of cancer in blood cell images is often vague, can overlap with other diagnoses, and can mimic many other benign abnormalities. These discrepancies cause considerable variability among medical personnel in the diagnosis of cancer. Automated detection of cancer from blood cell images at the level of expert medical personnel would not only have tremendous benefit in clinical settings, it would also be invaluable in delivery of health care to populations with inadequate access to diagnostic imaging specialists.

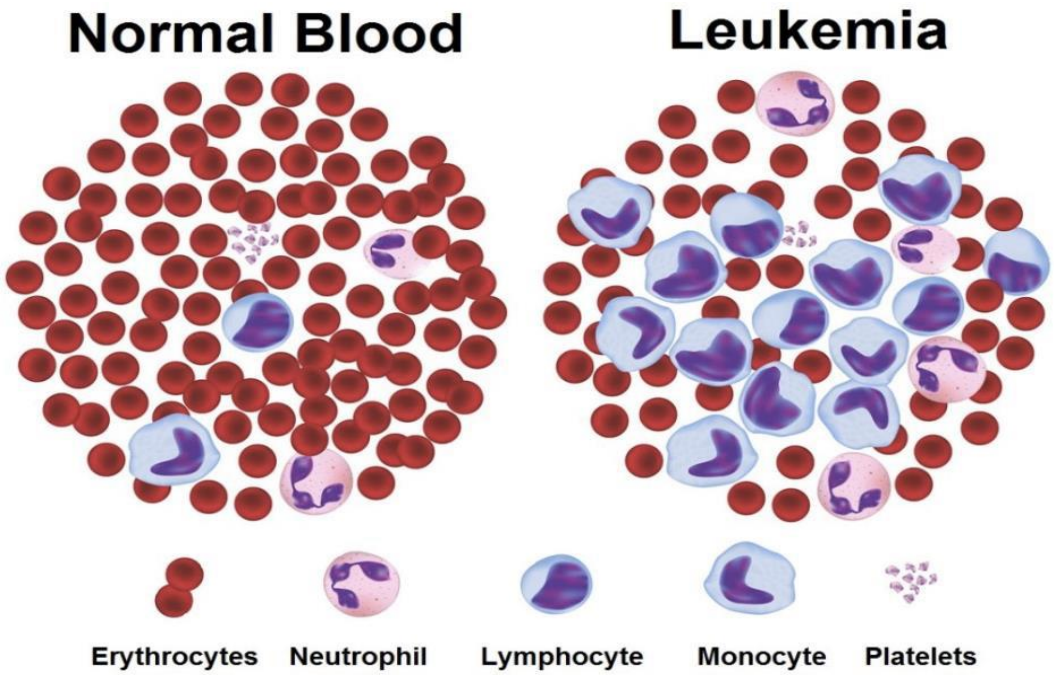


Figure1(b): Normal blood versus blood of affected patient (wisegeek.com)

1.3 SCOPE AND APPLICATION

We develop a system which detects cancer from blood cell images at a level exceeding practicing medical personnel. This technology can improve healthcare delivery and increase access to medical imaging expertise in parts of the world where access to skilled medical personnel is limited.

2. OBJECTIVES

The main objective of our project is to develop a system that detects whether the blood cell is cancerous or not.

3. LITERATURE REVIEW

Various techniques have been developed by researchers to detect leukemia. One of the most used technique is Convolution Neural Network (CNN) It is based on computer vision in recent years. The common algorithm for this approach consists of several rigid steps: image pre-processing, clustering, morphological filtering, segmentation, feature selection or extraction, classification, and evaluation.

[1]3.1 EXISTING METHODS FOR DIAGNOSIS

- Medical history and physical examination: The record of present symptoms, and problems a person has had in the past. The medical history of a person's family also helps in diagnose leukemia.
- Complete blood count (CBC): Blood is taken and checked under the microscope for the number of RBCs, WBCs and platelets.
- Bone marrow aspiration: Bone marrow is removed with the help of a needle from breastbone. The removed sample is observed under a microscope to look for abnormal cells.
- Cytogenetic analysis**:** Cytogenetic test takes blood or bone marrow to help identify individual chromosomes. It shows abnormalities in chromosomes, which help to diagnosis and identify the type of leukemia. Results are usually available within 3 weeks.
- Immunohistochemistry**:** Blood sample of cells are treated with special antibodies in immunohistochemistry. Under the microscope the change in color can be seen. It helps in determining the types of cells that are present.

3.2 CNN TECHNIQUES

[2] have proposed the segmentation method using color-based clustering to obtain nucleus region and cytoplasm area from stained blood smear images. SVM classifiers are applied with relevant features and gain satisfactory results.

[3] have proposed an automatic detection of white blood cells (WBCs) from peripheral blood images and classification of five types of WBCs: eosinophil, basophil, neutrophil, monocyte, and lymphocyte. Eosinophil and basophil from other WBCs are first classified by SVM with a granularity feature. Other three types are then recognized using convolutional neural network to extract features, and random forest uses these features to classify those WBCs.

[4] have introduced deep learning as a technique to improve the objectivity and efficiency of histopathologic slide analysis. Convolutional neural networks are trained in two experiments which are prostate cancer identification in biopsy specimens. They show that this system holds great promise to reduce the workload of pathologists with increasing objectivity of diagnosis.

[5] have proposed the method to categorize the two types of leukemia, ALL and acute myeloid leukemia (AML). The segmentation of blood cells is performed using contextual color and texture information to identify nucleus and cytoplasm region as well as to separate overlapped blood cells. The morphological, statistical, texture, size ratio, and eigen values features are extracted after segmentation to be used by various machine learning classifiers available in Weka.

[6] have used deep learning method based on a superpixel and convolutional neural network to detect the cytoplasm region in cervical cancer cell segmentation.

[7] have implemented a CNN classifier to explore the feasibility of deep learning approach to identify lymphocytes and ALL subtypes, and this approach is benchmarked against a dominant approach of support vector machines (SVMs) applying handcrafted feature engineering. This shows a great potential for image classification with no requirement of multiple preprocessing steps from feature engineering.

[8] have presented fully automatic system able to recognize 17 classes of myelogenous leukemia from images of bone marrow aspirate. Cells are segmented using watershed algorithm combined with region-growing and edge detection techniques. 117 descriptive features have been generated and selected using linear SVM.

4. REQUIREMENT ANALYSIS

4.1 FUNCTIONAL REQUIREMENT

4.1.1 SOFTWARE REQUIREMENT

Python

Data science is most effectively studied technology using python because python has many in built libraries that supports its use. Scikit-learn is a simple yet effective tool used for data analysis and is reusable in various context. Classification a type of supervised learning is also supported with the use of Scikit-learn and since from our collected data set and its compatibility, we choosed to implement KNN to achieve our required model and used it as our backend.

4.1.2 HARDWARE REQUIREMENT

Since, we will be taking the microscopic cell images as input that a desktop computer can read we will not require any additional hardware other than a desktop computer.

4.2 NON FUNCTIONAL REQUIREMENTS

1. Accuracy

The software which we are making should meet all the specification and should give the accurate results.

2. User interface

The system will maintain an easy accessibility and user friendly interfacing across all functionality and for all users.

3. Scalability

The system will be able to maintain its scale based on the number of users using the system.

4.Portability

The system should run on different device without any malfunctioning.

5.Maintainability

The system should be easy to maintain and there should be a separate interfacing of each block of code.

4.3 FEASIBILITY STUDY

The feasibility study activity involves the analysis of the problem and collection of all relevant information relating to the product. The main aim of the feasibility study

activity is to determine whether it would be financially and technically feasible to develop the product.

1.Technical Feasibility

This is concerned with specifying equipment and software that will successfully satisfy the requirements. Technical feasibility assesses the current resources (such as hardware and software) and technology, which are required to accomplish user requirements in the software within the allocated time and budget. For our system the required software and hardware is easily available to us. So we can say that it is technically feasible.

2.Economic Feasibility

Economic feasibility determines whether the required software is capable of generating financial gains for an organization. It involves the cost incurred on the software development team, estimated cost of hardware and software, cost of performing feasibility study, and so on. As the required software and hardware is not so expensive and easily available. So we can say the product is economically feasible.

3.Legal Feasibility

Legal feasibility is mainly related to human organizational and political aspects. The political and the constitution of the country has been taken into consideration while performing the legal feasibility. It won't share the personal information of the patient with anyone and keeps the privacy and etc. As the product does not violets against the law of our country. So it can also be legally feasible.

5. SYSTEM DESIGN AND ARCHITECTURE

5.1BLOCK DIAGRAM

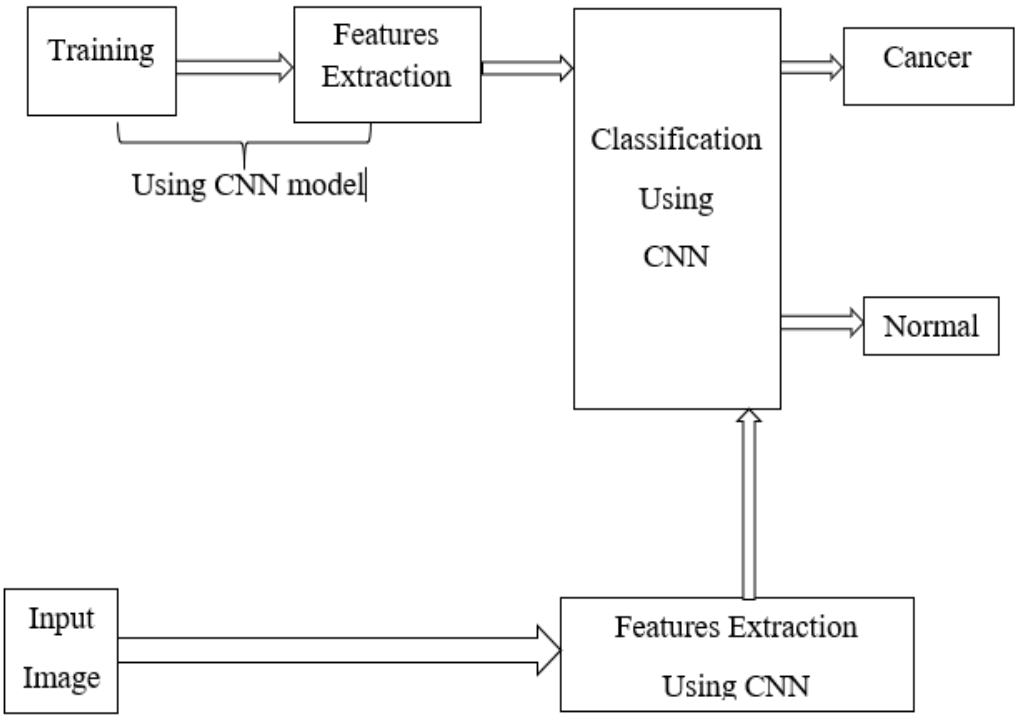


Figure5(a): Block Diagram of System

5.2 USE CASE

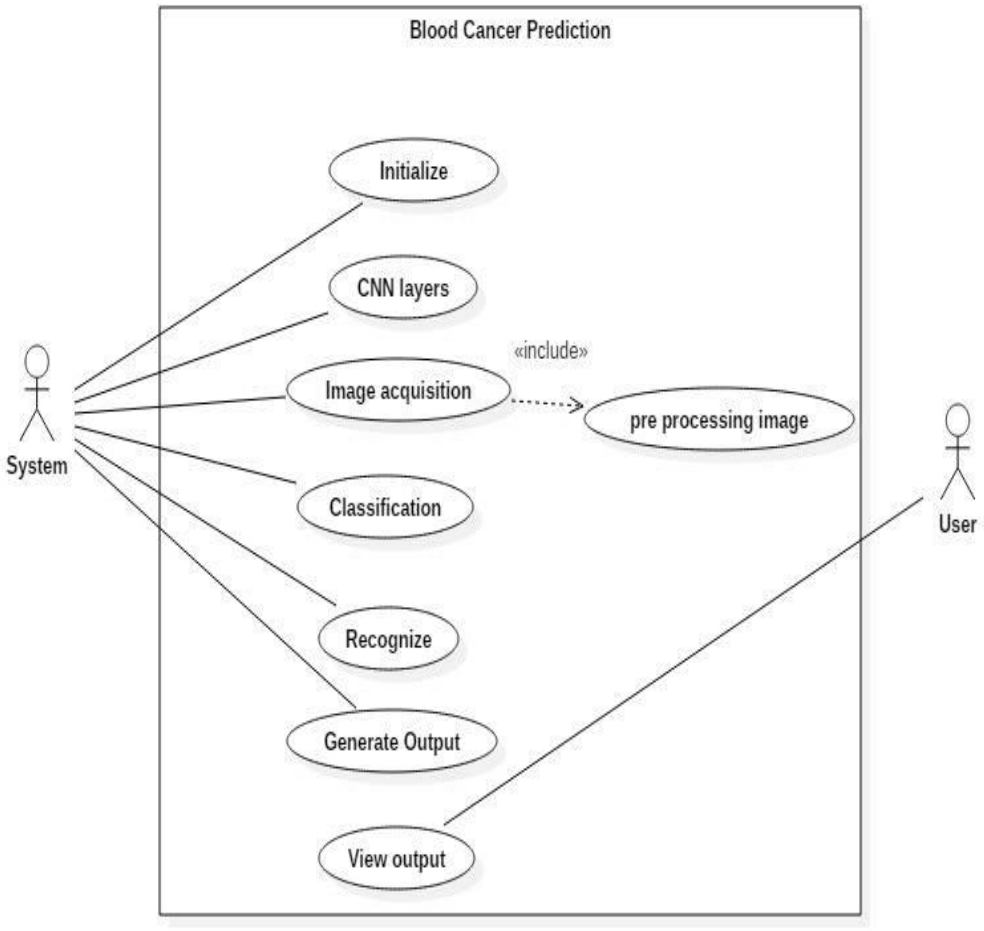


Figure5(b): Use Case Diagram

5.3 DFD

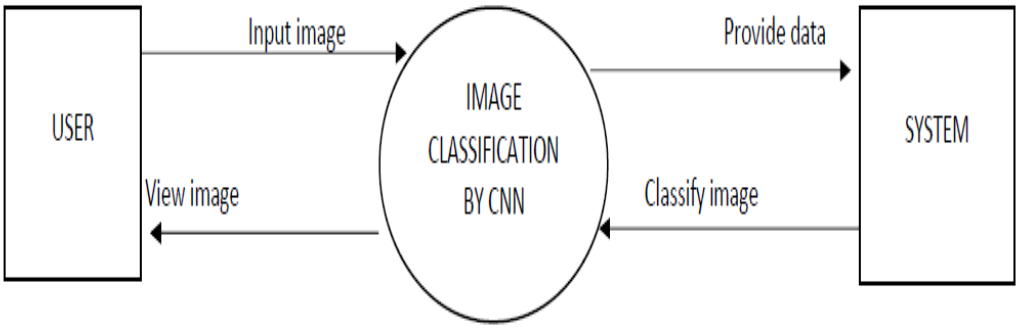


Figure5(c):DFD level 0

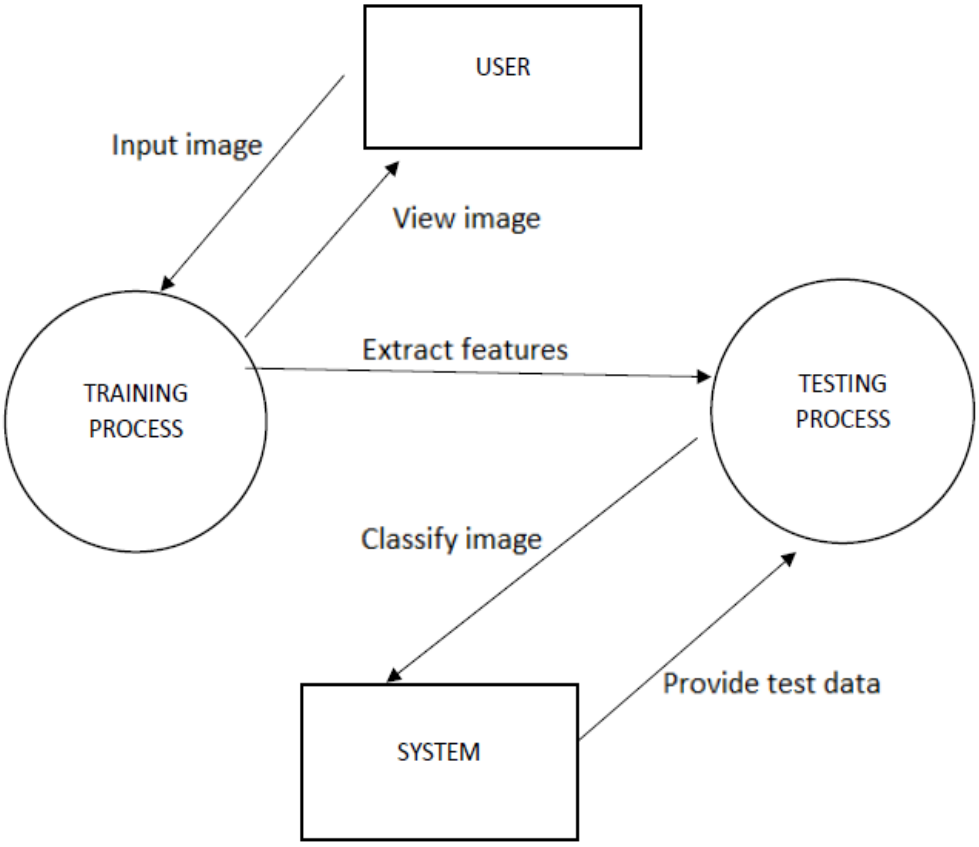


Figure5(d):DFD Level 1

5.4 SEQUENCE DIAGRAM

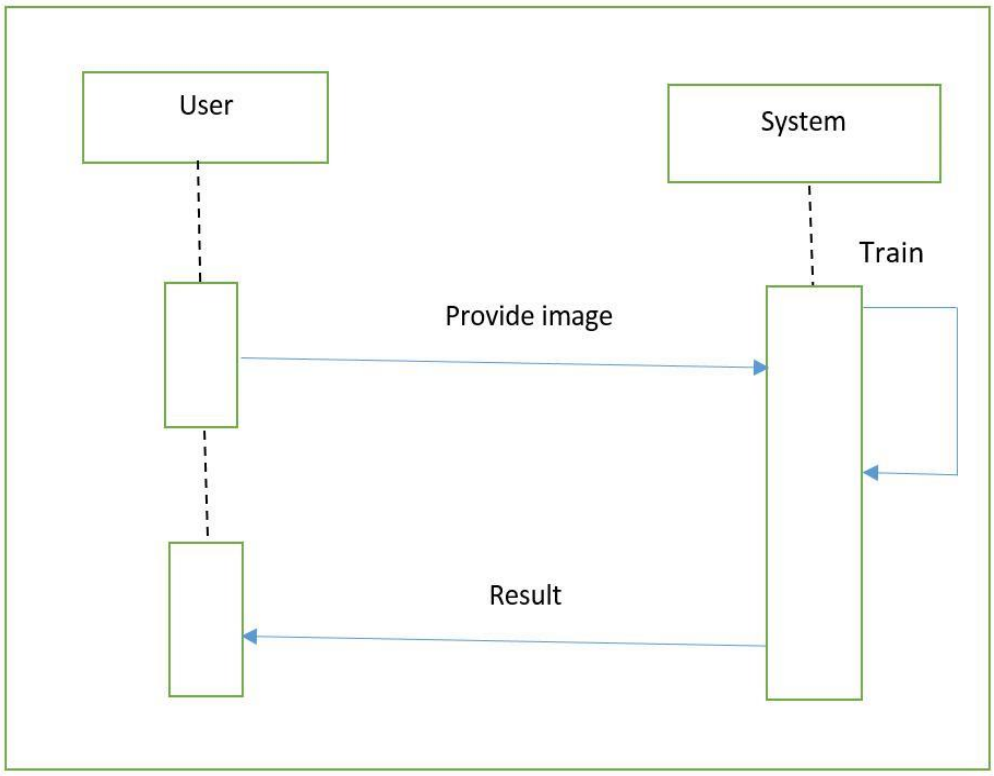


Figure5(e): Sequence Diagram

6. METHODOLOGY

6.1 SOFTWARE DEVELOPMENT APPROACH

Here for this project we will be following the Waterfall Model for the software development. We will follow this model because it is very simple and easy to use and understand. Since our project is small and requirements are properly defined and understood, this model will properly help us to complete our project.

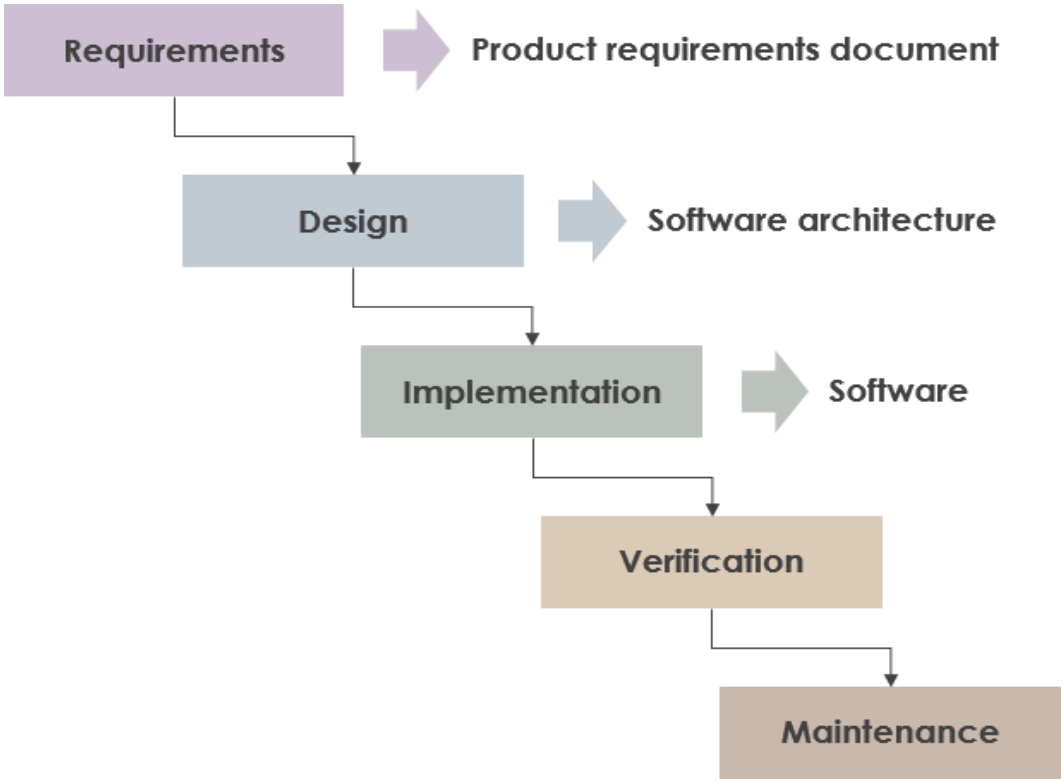


Figure6(a): Overview of Waterfall Model(medium.com)

6.2 DATASET

Image used in this project were obtained from Kaggle dataset which is a public dataset available online [9]. This dataset was divided into 2 classes. There was total 4961 training images where 2483 images were from healthy patients and 2478 images were from patients affected with blood cancer. We tested the model with total 1240 images 620 from each class. These images had resolution of 320*240.

Table 6(a): The number of training and test images

	Training Set	Testing Set
Normal Cell	2483	620
Cancerous Cell	2478	620
Total	4961	1240

6.3 CNN OVER OTHER ALGORITHMS

There are a lot of algorithms that people used for image classification before CNN became popular. People used to create features from images and then feed those features into some classification algorithm like SVM. Some algorithm also used the pixel level values of images as a feature vector too. To give an example, you could train a SVM with 784 features where each feature is the pixel value for a 28x28 image.

CNNs can be thought of automatic feature extractors from the image. While if we use a algorithm with pixel vector we lose a lot of spatial interaction between pixels, a CNN effectively uses adjacent pixel information to effectively down sample the image first by convolution and then uses a prediction layer at the end.

This concept was first presented by Yann le cun in 1998 for digit classification where he used a single convolution layer. It was later popularized by Alex net in 2012 which used multiple convolution layers to achieve state of the art on image net. Thus, making them an algorithm of choice for image classification challenges henceforth.

6.4 WORKING OF CNN

We have implemented CNN for the feature extraction and classification of the blood samples.

A CNN is a multilayered neural network with a special architecture to detect complex features in data. CNNs have been used in image recognition, powering vision in robots, text in images and for self-driving vehicles.

The CNN consist layer of neurons and it is optimized for two-dimensional pattern recognition. CNN has three types of layer namely convolutional layer, pooling layer and fully connected layer.

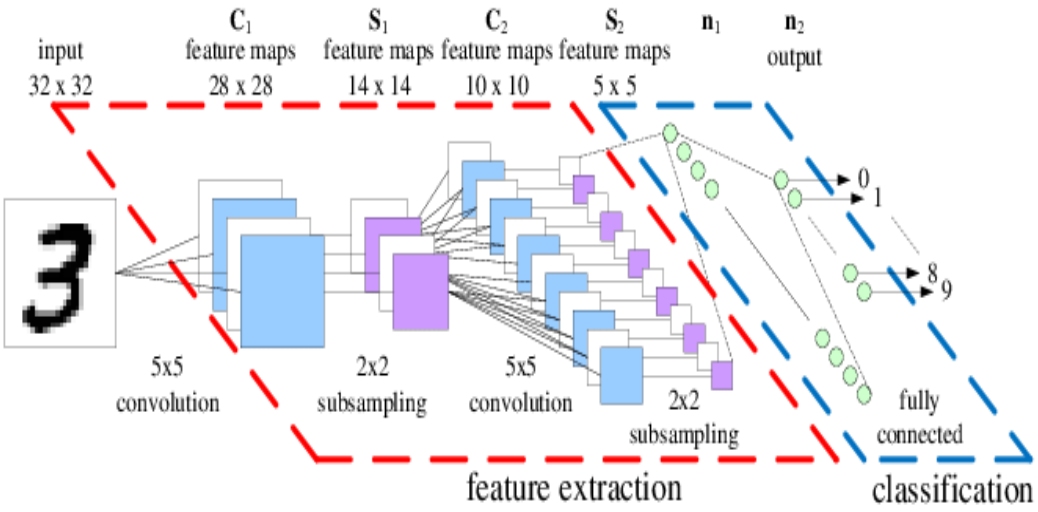


Figure6(b):Sample Convolution Neural Network Architecture(medium.com)

Our network consists of 11 layers excluding the input layer. The input layer takes in a RGB color image where each color channel is processed separately.

The first 6 layers of convolution network are convolution layer. First 2 convolution layer applies 16 of 3*3 filters to an image in the layer. The other two layer applies 32 of 3*3 filters to an image. And the last 2 layers of convolution applies 64 of 3*3 filters to an image. The nonlinear transformation sublayer employs the ReLU activation function. The max pooling sublayer applies a 2*2 filter to the image which results in reducing the image size to its half. At this point, convolution network extracts 64 features, each represented by a 32*32 array for each color channel.

The eighth layer is the flatten layer. The flatten layer transforms a multidimensional array into one-dimensional array by simply concatenating the entries of the multidimensional array together. The output of this flatten layer is a one-dimensional array of size 4800. The ninth layer is the fully connected ANN with the ReLU activation function that maps 4800 input values to the 64 output values. The tenth layer is the dropout layer. 50% of the input values coming to the layer are dropped to zero to reduce the problem of overfitting. The eleventh and the final layer is a fully connected ANN with the sigmoid activation function that maps 64 input values to 2 class labels.
First, we train convolution network using the data in training set to find appropriated filters' weights in the three convolutional sublayers and the weights that yield minimum

error in the two fully connected layers. Next, we evaluate convolution network using the data in the validation set to obtain validation error and cross-entropy loss. We repeat the training of convolution network in this same procedure until we complete 10 epochs. Last, we evaluate the performance of convolution network using data in the test set.

6.5 CLASSIFICATION

Neural networks are used in the automatic detection of cancer in blood samples. Neural network is chosen as a classification tool due to its well-known technique as a successful classifier for many real applications. The training and validation processes are among the important steps in developing an accurate process model using CNNs. The dataset for training and validation processes consists of two parts; the training features set which are used to train the CNN model; whilst a testing features sets are used to verify the accuracy of the trained using the feed- forward back propagation network. In the training part, connection weights were always updated until they reached the defined iteration Number or suitable error. Neural networks are used in the automatic detection of cancer in blood samples. Neural network is chosen as a classification tool due to its well-known technique as a successful classifier for many real applications. The training and validation processes are among the important steps in developing an accurate process model using CNNs.

7. RESULTS AND ANALYSIS

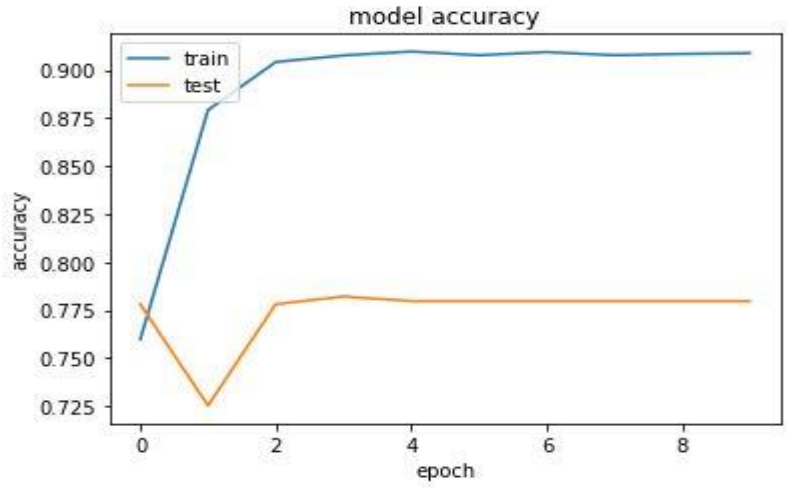
The final output of our project is to detect the cancer accurately with the help of iterations obtained, loss and accuracy graph and the confusion matrix.

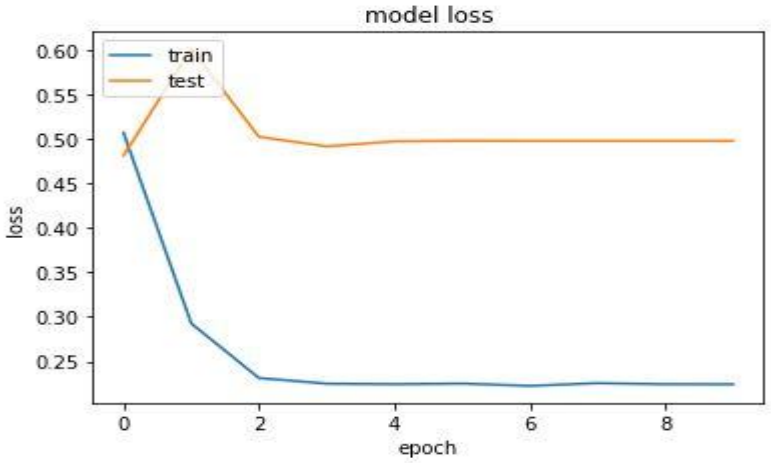
We have performed twenty series of iteration from which we can clearly observe that the loss is decreasing with each iteration. Loss is about how much right the model is. So we wanted to minimize the loss function and as a result our model has perfectly declined the loss value straight from starting point and at every iteration we get closer to minimum.

```
Epoch 00008: ReduceLROnPlateau reducing learning rate to 4.999999858590343e-11.

Epoch 00008: val_acc did not improve from 0.78205
Epoch 9/10
5216/5216 [==============================] - 7s 1ms/step - loss: 0.2243 - acc: 0.9085 - val_loss: 0.4974 - val_acc: 0.7796

Epoch 00009: ReduceLROnPlateau reducing learning rate to 4.999999719812465e-12.

Epoch 00009: val_acc did not improve from 0.78205
Epoch 10/10
5216/5216 [==============================] - 7s 1ms/step - loss: 0.2242 - acc: 0.9088 - val_loss: 0.4974 - val_acc: 0.7796

Epoch 00010: ReduceLROnPlateau reducing learning rate to 4.999999546340118e-13.

Epoch 00010: val_acc did not improve from 0.78205
```

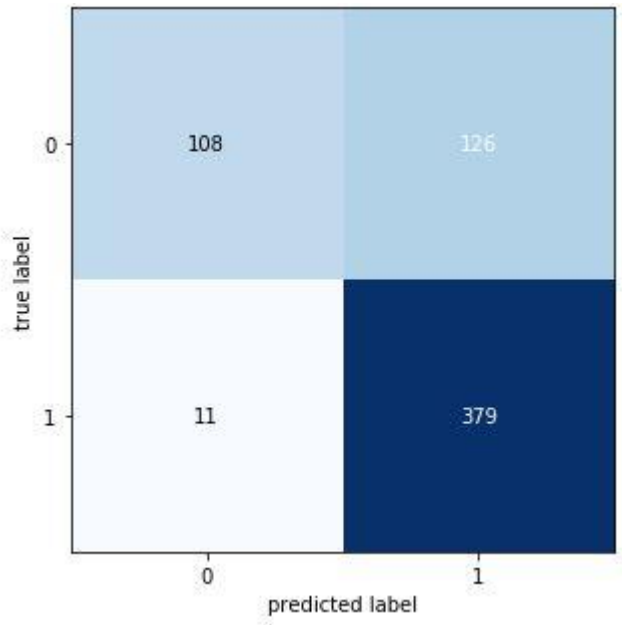
Next we performed loss and accuracy curve for the best result of our model. These learning curves (loss and accuracy curve) shows the performance of our model on training and validation set as a function of number of training iterations.

We have loss curve which is decreasing with each iteration which shows that loss is minimizing giving the best result. On the other hand, we have performed accuracy curve which is increasing with each iteration that means our model is getting better and better at learning.





Finally, we evaluated confusion matrix to describe the performance of our classification model. We then calculated precision, accuracy and recall value from the obtained confusion matrix to measure our model accurate performance.



Here the 0 belongs to class of people not having cancer and 1 belongs to class of people having cancer. The confusion matrix consists of True positive, True negative, False positive and False positive values according to which different parameters are calculated which is shown in figure below:

```
In [16]: #PRECISION = (TP/(TP+FP))
         379/(379+126)

Out[16]: 0.7504950495049505

In [17]: #RECALL = (TP/(TP+FN))
         379 / (379 + 11)

Out[17]: 0.9717948717948718

In [18]: #ACCURACY = (TP+TN)/(TP+TN+FP+FN)
         (379+108)/(379+108+126+11)

Out[18]: 0.780448717948718
```

Here, the recall is most significant quantity even more than accuracy and precision. Since we are having unequal number of people in both the classes, therefore we can't take accuracy as an alone metric to calculate model accuracy. Also, we have to minimize the false negative which is in the denominator of recall increasing the value for recall.

False negative has to be intuitively minimized because falsely diagnosing a patient of Cancer as not having Cancer is much larger deal than falsely diagnosing a healthy person as a Cancer patient which is our major concern. That is why we are making this model to reduce the mistakes done by doctors accidentally.

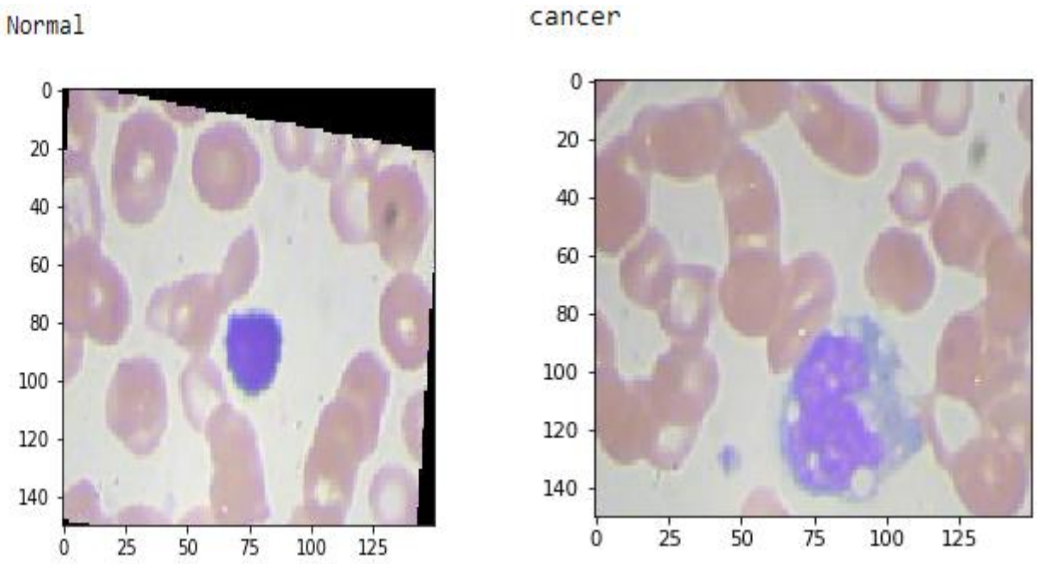


Figure7(a):Output Results

9. TOTAL COST

In the project of detection of blood cancer using image CNN, we used microscopic images of the blood cells taken from the internet, hence we do not require any extra hardware for the image acquisition purpose. We used a desktop computer only and no other hardware device is required. This can be easily implemented and hence won't require any extra cost for the project.

10. CONCLUSION

In this work, we present a deep learning approach to recognize cancerous cell defined by WHO classification. We implement a CNN, which directly takes raw images and automatically discovers useful features through a series of multilayer architecture. Using the model we were able to gain accuracy of 78.04 %. Instead of traditional approach of image segmentation CNN was successful in giving more accuracy than those prior approach.

11. LIMITATIONS AND FUTURE WORK

Still there are many flaws in our system which can be overcome in future.

11.1 LIMITATIONS

Our current system works on command line approach. Due to this, all the users may not find the system an easy software to run and work. Also the system is less interactive hence, user may require lot of time to operate and get the result.

11.2 FUTURE WORK

1. This system can be further extended to detect subtypes of cancer(ALL,AML,CML,CLL) found in blood.
2. This system can be extended to implement in android/iOS.
3. The system will find the stages of blood cancer.

REFERENCES

[1]Chaitali R., and Jyoti Rangole, "Detection of Leukemia in microscopic images using image processing": International Conference on Communications and Signal Processing (ICCSP),2014.

[2]S. Mohapatra, D. Patra, and S. Satpathy, "Unsupervised blood microscopic image segmentation and unsupervised blood microscopic image segmentation and leukemia detection using color based clustering," International Journal of Computer Information Systems and Industrial Management Applications, vol. 4, pp. 477–485, 2012.

[3]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, 2016.

[4]G. Litjens, C. I. Sánchez, N. Timofeeva et al., "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," Scientific Reports, vol. 6, no. 1, 2016.

[5]C. Reta, L. A. Robles, J. A. Gonzalez, R. Diaz, and J. S. Guichard, "Segmentation of bone marrow cell images for morphological classification of acute leukemia," in Proceedings of the 23rd International FLAIRS Conference, Daytona Beach, FL, USA, May 2010.

[6]Y. Song, L. Zhang, S. Chen et al., "A deep learning based framework for accurate segmentation of cervical cytoplasm and nuclei," in Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Chicago, IL, USA, August 2014.

[7]department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani 94000, Thailand. Published 2 June 2019.

[8]S. Osowski and T. Markiewicz, "Support vector machine for recognition of white blood cells in leukemia," in Kernel Methods in Bioengineering, Signal and Image Processing, pp. 93–123, Idea Group Inc, Calgary, Canada, 2006.

[9] https://www.kaggle.com/paultimothymooney/blood-cells(2019/05/10)

[10] K. He and J. Su, "Convolutional neural networks at constrained time cost," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, June 2015

Table of Contents