Leukemia Detection and Classification

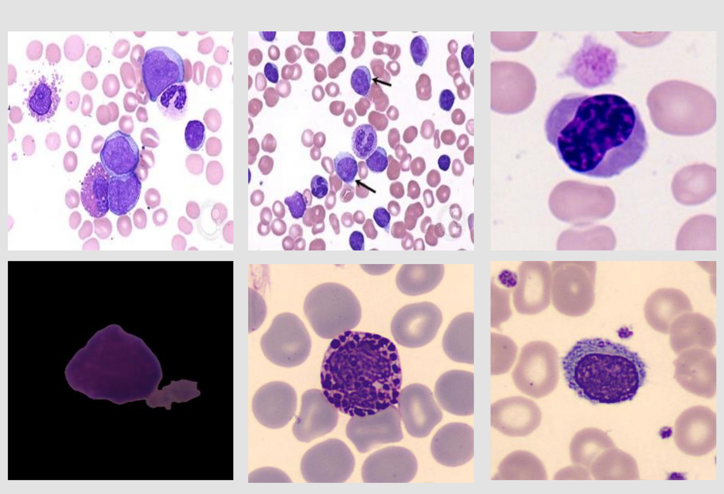
Esraa Ramdan, Nada Samir, Samar SalahElDin Ghoneim, Toka Mohamed, Somaya Yasser
Dr. Ahmed Fraouk, Dr. Hanaa Mobarz, TA. Sara ElNady

Abstract

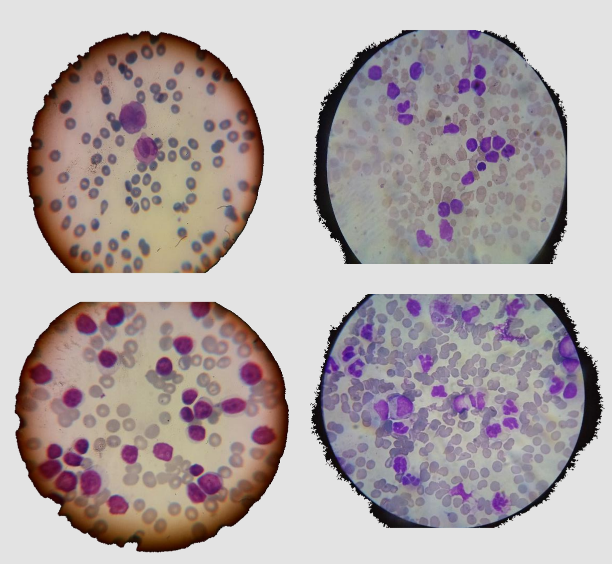
Leukemia is a blood cancer that can be fatal in most cases. Detection and classification of Leukemia into which type the patient has can take a lot of time and resources for such process to be done. There are multiple stages for the process to know whether the patient is affected or not and which type do they have to determine after the prognosis and which treatment is needed.
We used two datasets one that had 5 publicly available data sources for each Leukemia type; Acute lymphocytic Leukemia, Chronic lymphocytic Leukemia, Acute Myeloid Leukemia and Chronic Myeloid Leukemia, and normal blood cells. The second had 4 leukemia types; Acute lymphocytic Leukemia, Chronic lymphocytic Leukemia, Acute Myeloid Leukemia and Chronic Myeloid Leukemia. Data Augmentation was then applied when needed. Finally, we want to show the variety of performance between CNN architecture models and other well-known machine learning algorithms.

Introduction

Leukemia is a cancer in the blood that affects people of all ages and of any gender. It can lead to destroying the immune system of the human body and makes the bone marrow produce excess amount of abnormal WBCs, which do not function properly. Leukemia can be classified as either myeloid or lymphoid. The two main types are divided into two subtypes: acute and chronic. Therefore, Leukemia types are: Acute lymphoid Leukemia, Chronic lymphoid Leukemia, Acute Myeloid Leukemia and Chronic Myeloid Leukemia.
The traditional way to diagnose leukemia requires numerous amounts of steps which require a lot of time and resources. Therefore, we are taking a machine learning/deep learning approach to do these steps in less time and less energy taken. Given a dataset of microscopic blood images, we want to train and build several models than can detect Leukemia in these images and classify it as well.
Through our study, we are working on Detection of Leukemia (whether the patient is affected or not) and Classification into its types if affected (ALL, AML, CML and CLL). We are approaching the problem by applying machine learning and deep learning on our data. We are going to use SVM algorithm and CNN algorithm.
By using the model that is going to be created, specialists will save both time and effort for a primary detection and classification of the patient status regarding Leukemia. We are taking both an image processing approach and a bioinformatics one then comparing the results of both of them.



First dataset

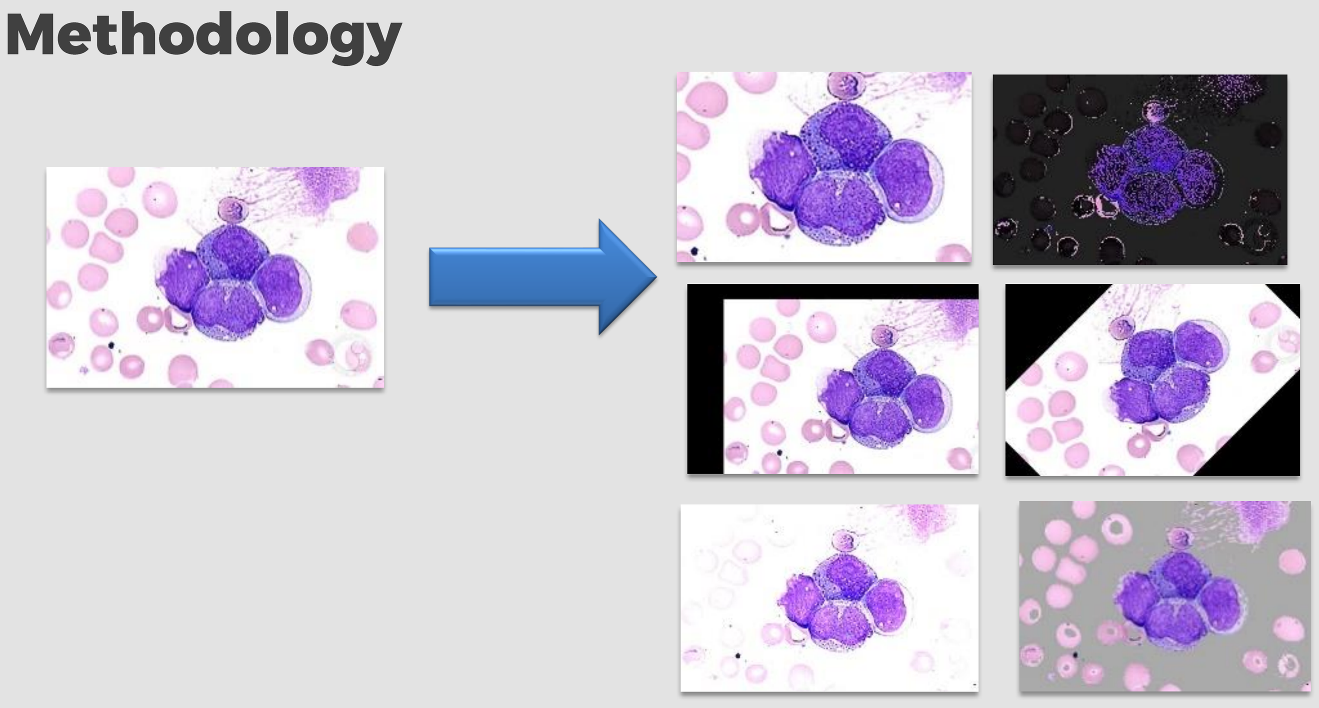


Second dataset

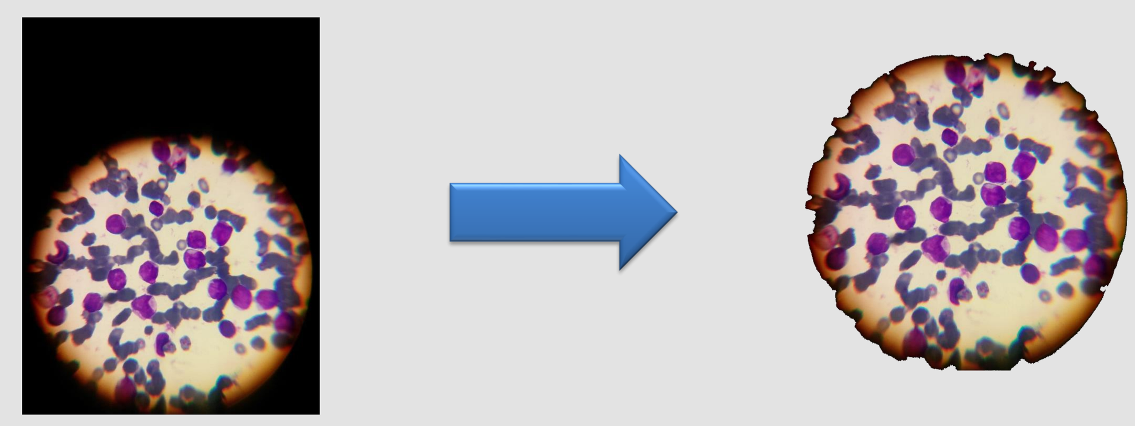
Methodology

Firstly, We have applied data augmentation on the first dataset only due to the huge different in the number of images between CLL and CML types and the rest of the types (AML, ALL, and Normal). Unlike the second dataset where the number of images were closer to each other. We used several ways to augment the data, such as; Flipping, Rotation, Translation, Scaling, Brightness, Saturation, Changing Color, and Cropping.
Secondly, We have applied some preprocessing on the image before it being trained on the models we used. We resized all the images to be of the same dimensions, they ended up to be 124x124 pixels. Then, converting them to the same file extension, jpg for the first datasets. Whereas for the second dataset, it already had the jpg extension. For the new dataset there was a lot of noise around the samples, so we removed it. For the deep learning models these steps were followed by the change of the images from RGBA to RGB for both datasets. However, for the machine learning models the preprocessing was as follows: Converting the images into grey scale, Feature Extraction using multiple methods (Edge detection using Gaussian blur and Canny Detection, Morphological transformation (divide, threshold, getStructure, etc.) and Image Contouring.)
We have performed Machine Learning and Deep Learning on both of our datasets. For machine learning we used SVM, Logistic Regression and Random Forest. For deep learning we used CNN and known architectures like VGG16, AlexNet, MobileNet, and ResNet50.

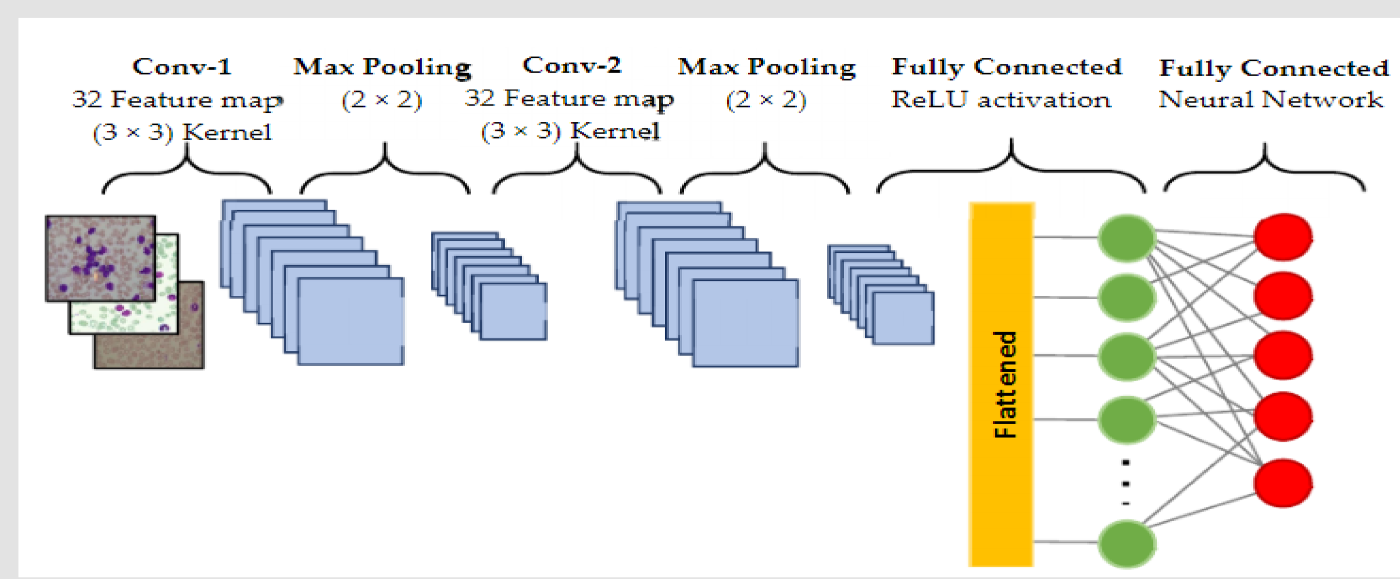
Methodology



Augmentation



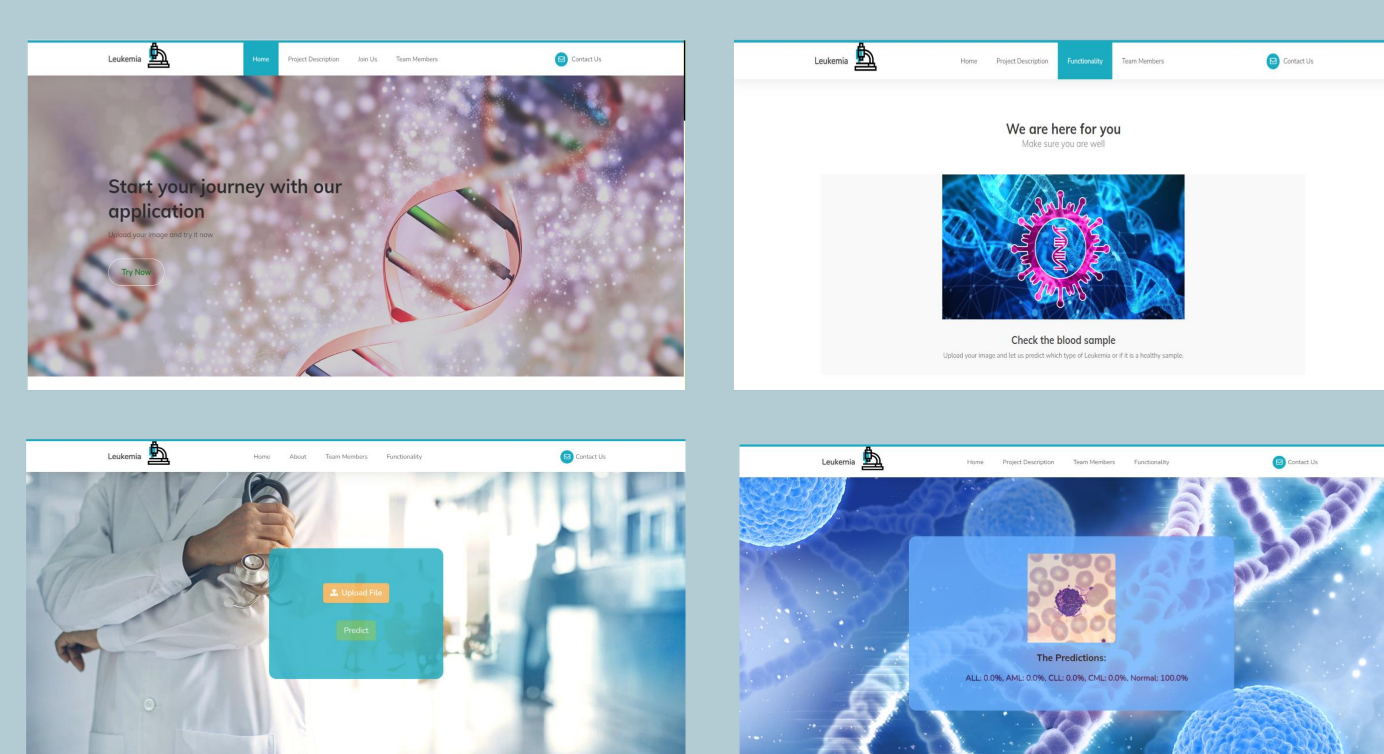
Preprocessing



CNN model architecture

Primary Design

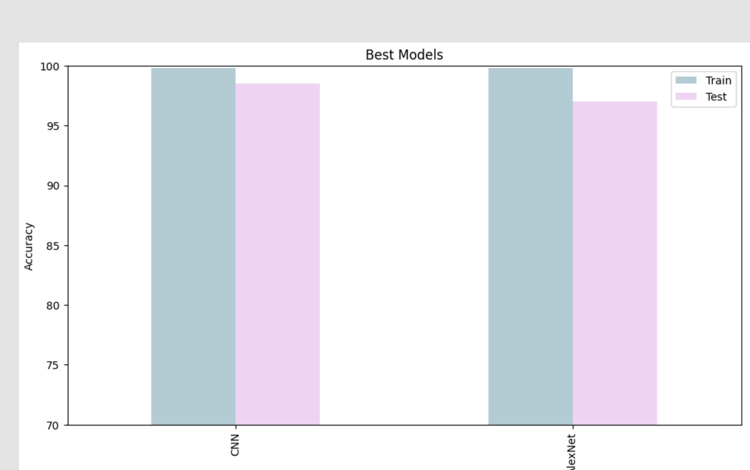
We built a user friendly website where the user can upload images of the samples needed for classification using our model.



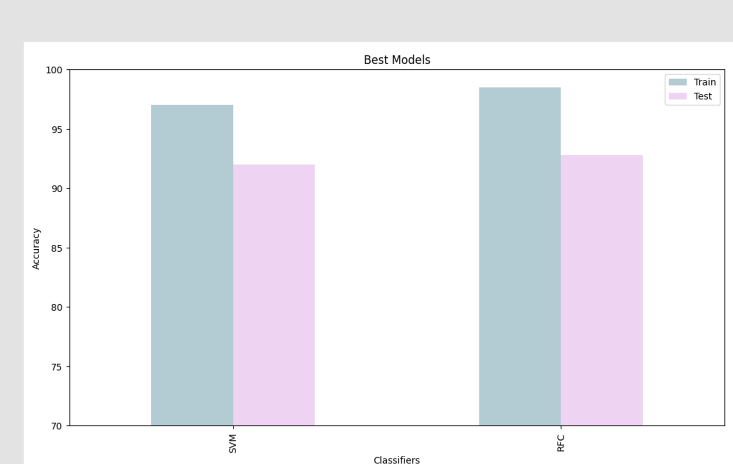
Obtained Pictures from Our website

Conclusion

After using multiple models and architectures, the best architecture for deep learning was using our **CNN model** with accuracy of **98.5%** followed by **AlexNet** with accuracy of **97%**. However, machine learning produced accuracies with **Random forest** of **92.8%** followed by **SVM** with accuracy of **92%**. Both were using our old dataset.



Train/Test Accuracies For Deep Learning



Train/Test Accuracies For Machine Learning