Childhood Leukemia Cell Classification

Name: Manita Ngarmpaiboonsombat, **Student ID:** 01940831

Email: manita_ngarmpaibooonsombat@student.uml.edu

Department: Computer Science, **Level:** Graduate

Abstract

Leukemia is one of the most common cancers in children. This project is focusing only on the Acute lymphoblastic leukemia (ALL) type which account for 75% of all leukemia in childhood. Currently there are no widely recommended screening tests for leukemia before it starts to cause symptoms. Therefore, classifying leukemia is a very important step for planning the treatment and predicting the treatment outcomes.

The purpose of this project aims to distinguish the abnormal white blood cells in the microscope images, focusing only on ALL leukemia type from the normal cells. The dataset comprises only white blood cell images with the type labels. This dataset is the binary classification which predicts the two output variables are Leukemia white blood cell (1) or normal white blood cell (0).

In this project, we used 3 classification algorithms, which are Naïve Bayes, logistic regression, and neural networks. The dataset is prepared by extracting the image features which are related to the pixel values and image circle detection and then selected them according to correlation. They were used as inputs for Naïve Bayes and logistic regression, while we did image classification with neural network. The result shows that the Naïve Bayes with categorical model fits better than others and generated the highest F1 score, which is 0.798, while the neural network has the lowest which is 0.76.

Introduction

Leukemia is one of the most common cancers in children, approximately one-third of all childhood cancers. About 3 out of 4 leukemias is Acute Lymphoblastic Leukemia (ALL) which is a cancer of the lymphoid line of blood cells. Currently, if the children don't have any signs or symptoms, there are no blood tests or other screening tests recommended by the doctors. Classifying leukemia is a very important step for planning the treatment as well as predicting the treatment outcomes but it is a very complicated process.

The purpose of this project aims to distinguish the abnormal white blood cells in the microscope images, focusing only on ALL leukemia type from the normal cells. Majority of children with leukemia will have too many white blood cells which will be blasted. Therefore, this model classification will help doctors suspect and identify a child with leukemia faster and easier before getting into other screening tests.

Details and Methods

- Since the images of the normal white blood cells and the immature leukemic blasts appear similar, therefore the process of identifying them is challenging.
- During their data acquisition, some staining noise and illumination errors remain in the microscopic images of these cells even though they have been mostly fixed.
- The dataset comprises of 7,271 Leukemia white blood cell images and 3,389 normal white blood cell images.

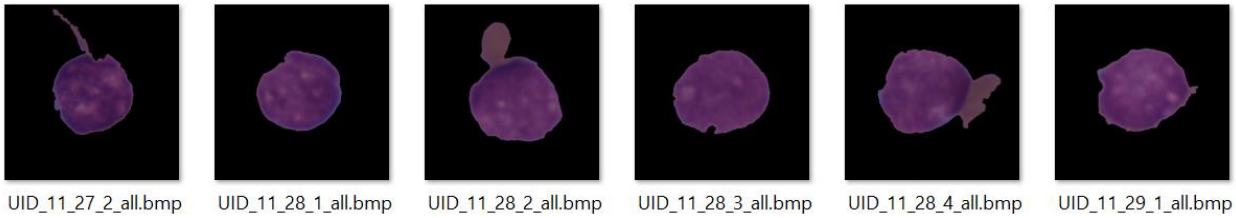

Figure 1 Examples of leukemia cells

- Set diagnosis = 1 for ALL (Acute lymphoblastic leukemia) and diagnosis = 0 for HEM (Normal white blood cell)
- Naïve Bayes and Logistic regression: prepared dataset by extracting features from the white blood cell images.
    - Pixel values
        - Max_NM_pixel: Normalized Pixel values in the range of 0-1 and take the maximum value
        - Mean_global: Calculated and subtracted the mean pixel value across color channels
        - SD_global: Calculated standard deviation across all color channels
    - Circle detection: Used function cv2.minEnclosingCircle() to detect circle which completely covers the object with minimum area and find the radius and center X, Y of the white blood cell. Since I think that leukemia white blood cells will be blasted, this should have an effect on the radius of the cell image.

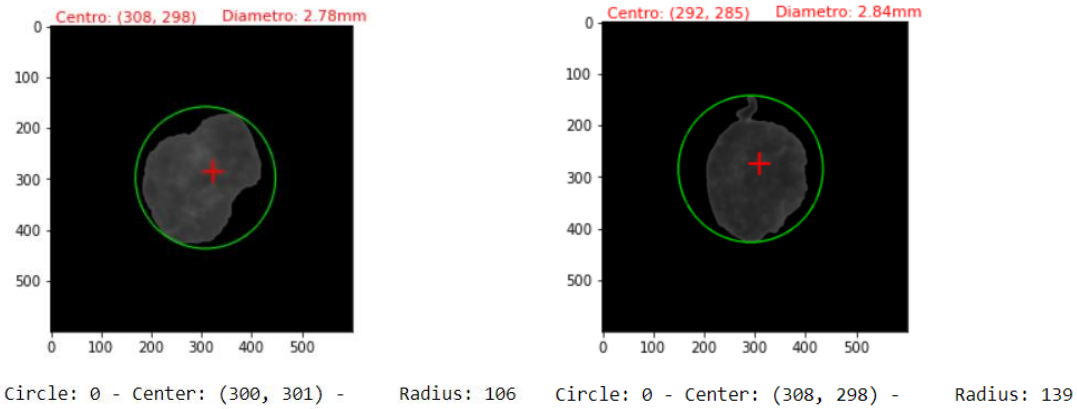

Figure 2 Minimum enclosing circle detection

- The table below shows an example of the dataset after extracting the features

	Max_NM_pixel	Mean_global	SD_global	center_x	center_y	radius	diagnosis
0	0.478431	13.924183	30.215250	225.0	241.0	128.0	1.0
1	0.498039	17.189129	33.770287	256.0	213.0	183.0	1.0
2	0.670588	19.350847	36.015213	207.0	203.0	180.0	1.0
3	0.505882	17.857874	34.366390	251.0	266.0	179.0	1.0
4	0.576471	11.770942	29.151751	229.0	229.0	113.0	1.0
5	0.580392	18.548267	37.723511	219.0	256.0	153.0	1.0
6	0.509804	12.616633	29.836288	236.0	225.0	157.0	1.0
7	0.490196	10.662043	26.687437	223.0	225.0	105.0	1.0
8	0.533333	16.989056	33.859528	308.0	338.0	29.0	1.0
9	0.525490	14.061645	31.758392	219.0	222.0	116.0	1.0

Figure 3 Table of example feature values

- Neural network: Ran image classification directly with image and label. I resized image to 128*128 and then stored images and labels in the np.array. This picture below shows the sample after shuffling the np.array.

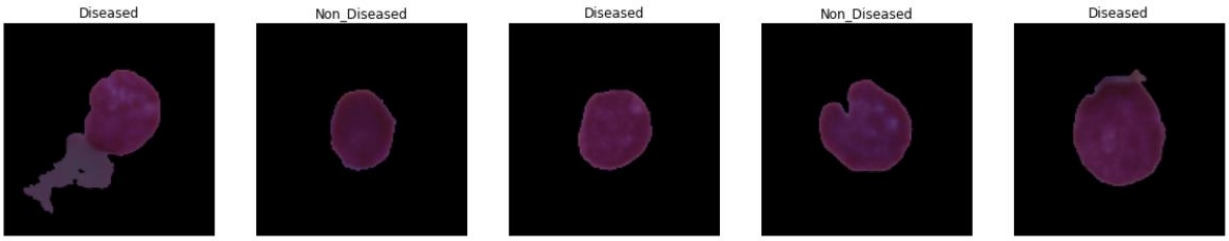

Figure 4 Examples of images and labels used for neural network training

Results

- Naïve Bayes
  1) Gaussian model, F1 score = 0.767
  2) Categorical model, F1 score = 0.798

  This picture below shows the correlation of all parameters. I selected 4 parameters which are Max_NM_pixel, Mean_global, SD_global and radius to run the model because center X and Y are barely correlated with the others.

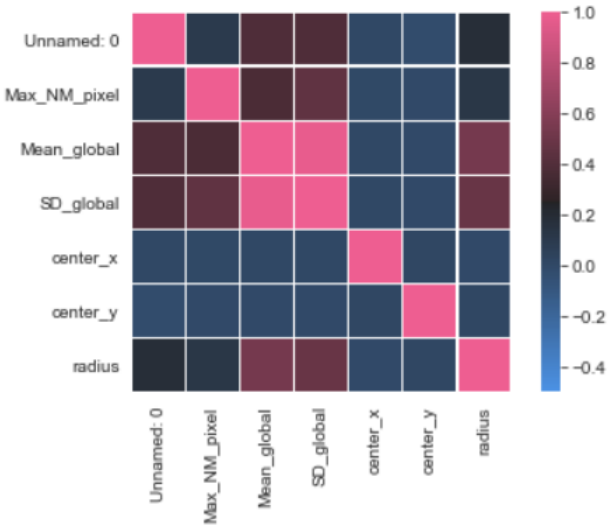

Figure 5 Correlation of the features

- Logistic Regression: F1 score = 0.78
    - Ran with the same dataset and parameters as Naïve Bayes.
    - This picture below shows the graph of loss function vs. epochs.

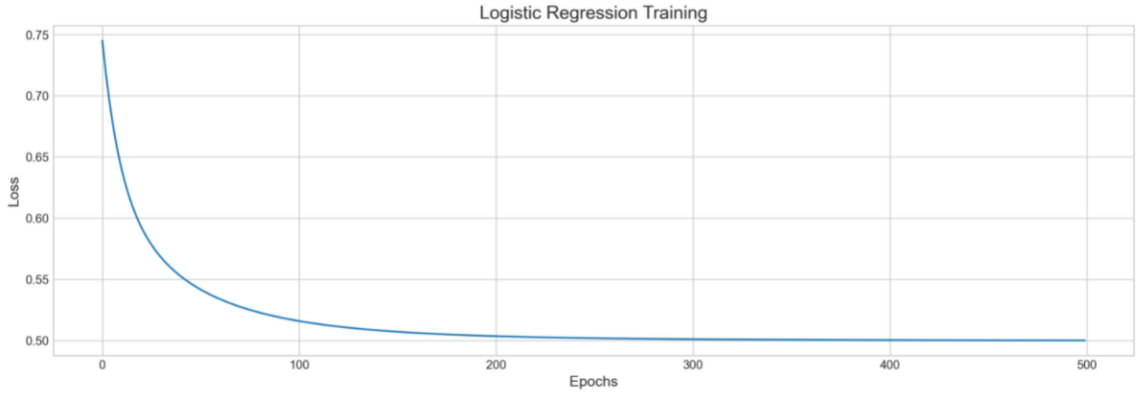

Figure 6 Loss function vs Epochs

- Neural network: F1 score = 0.76

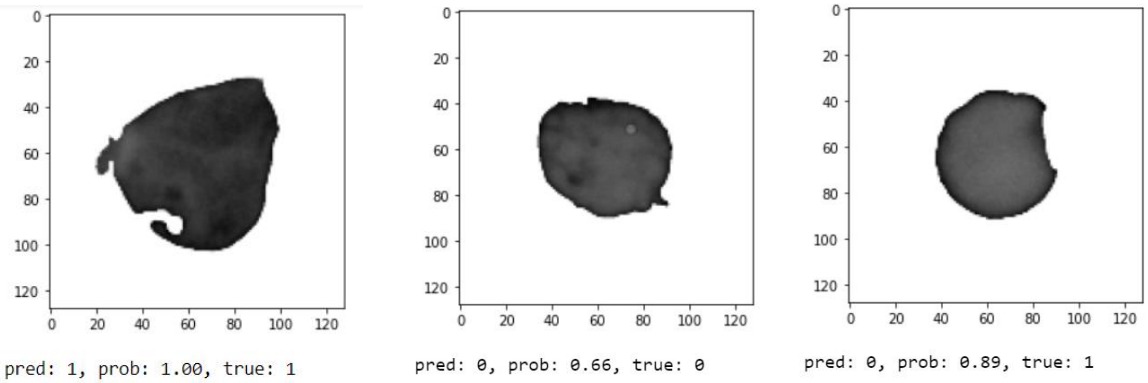

Figure 7 Examples of neural network results

Discussion

When comparing all three models, Naïve Bayes with categorical model fits better than others and generates the highest F1 score which is 0.798, while neural network has the lowest F1 score which is 0.76. The categorical Naïve Bayes is suitable for classification with discrete features that are categorically distributed.

In theory, a neural network is always better and more precise for binary classification problems than logistic regression and Naïve Bayes. However, in this experiment the neural network model is worse than others perhaps because there are nonlinearities involved and is more susceptible to overfitting than others. Another reason could be that the neural networks is more complex to train and requires a larger dataset for its optimization.

Conclusion

This project's dataset fits better with Naïve Bayes with categorical model which gives the F1 score of 0.798, while the Neural network generated the lowest F1 score, which is 0.76. The Naïve Bayes and Logistic regression are the powerful supervised algorithm used for binary classification problems. A neural network is more complex and must train the model carefully and it will produce better performance when you have sufficient training data.

References

[1] https://www.kaggle.com/andrewmvd/leukemia-classification
[2] https://www.cancer.org/cancer/leukemia-in-children.html