Tim Crnkovic

Classification of microscopic pathology images of Acute Lymphoid Leukemia and Multiple Myeloma using a Convolutional Neural Network.

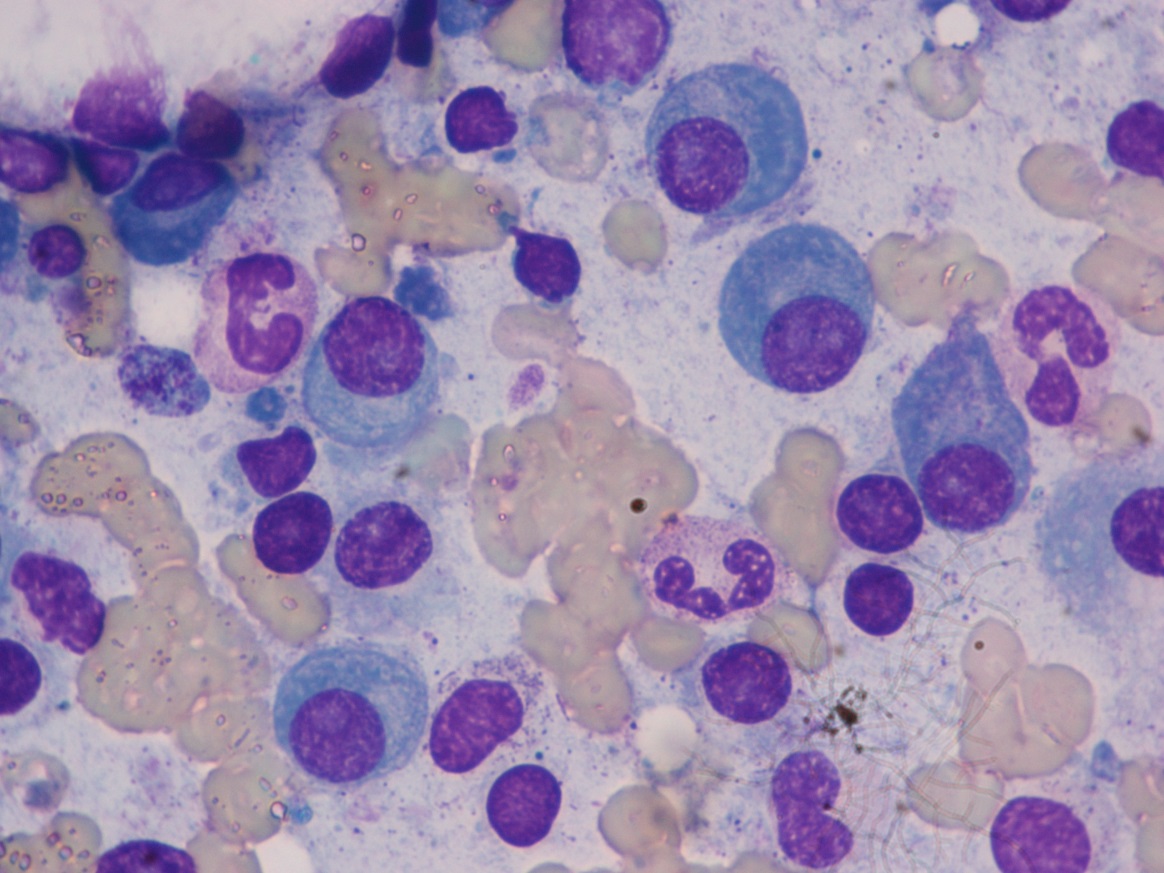
Introduction

The Cancer Imaging Archive (<https://www.cancerimagingarchive.net/> ) hosts de-identified collections of cancer images available for public use. From this archive, microscopic pathology images from patients diagnosed with B-lineage Acute Lymphoid Leukemia (B-ALL) and Multiple Myeloma (MM) were retrieved (<https://wiki.cancerimagingarchive.net/display/Public/SN-AM+Dataset%3A+White+Blood+cancer+dataset+of+B-ALL+and+MM+for+stain+normalization> ). These slides were stained using Jenner-Giemsa stain. This dataset consists of 29 images of B-ALL and 30 images of MM, from 59 different subjects.

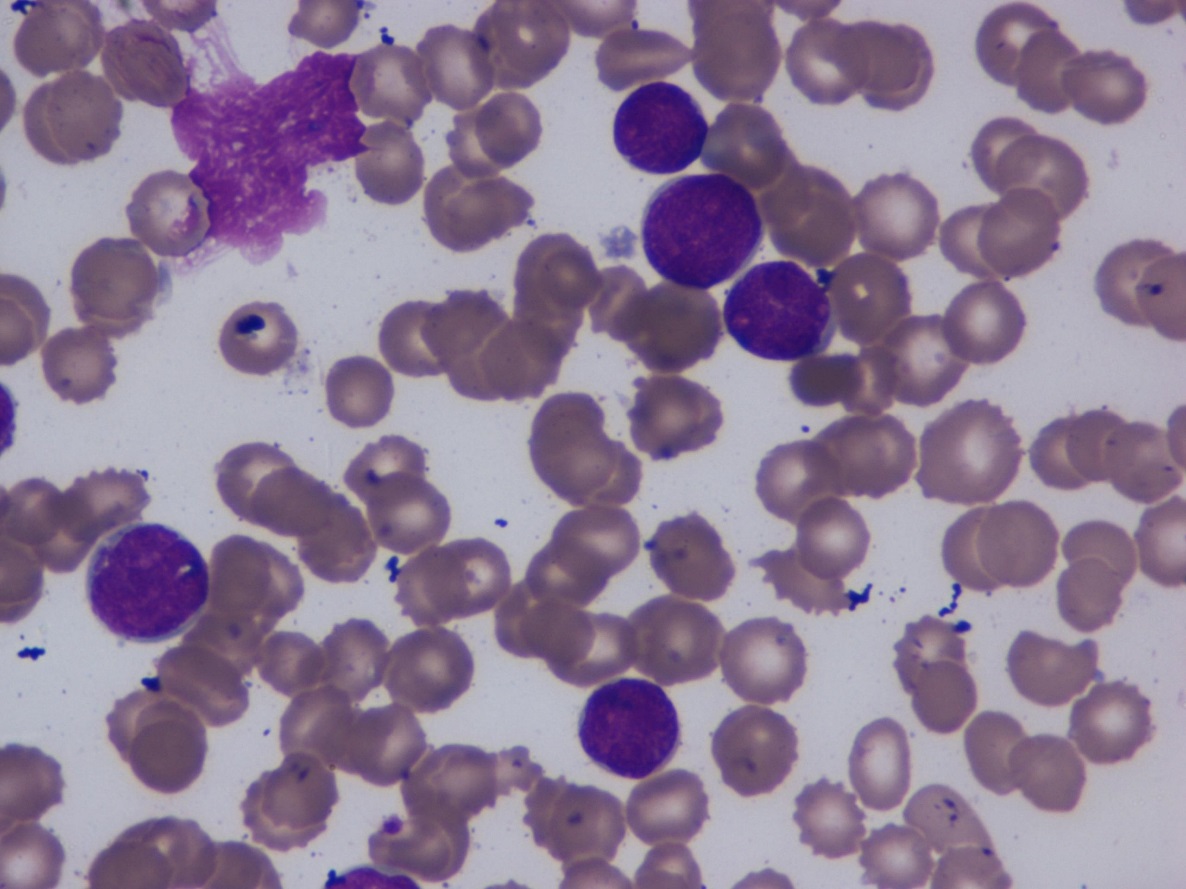
An additional 71 slides of MM from 5 different subjects were also obtained from the Cancer Imaging Archive (<https://wiki.cancerimagingarchive.net/display/Public/MiMM_SBILab+Dataset%3A+Microscopic+Images+of+Multiple+Myeloma> ). These microscopic pathology images were also stained using Jenner-Giemsa stain.

Furthermore, to augment the training/test data, each image was flipped on both axes and labelled. So the total sizes of the data sets are 58 images of ALL and 202 images of MM.

Examples of images of both types of malignancy are shown below.



*Example of a Multiple Myeloma stained pathology slide.*



*Example of an Acute Lymphoid Leukemia stained pathology slide.*

Methods and Results

Using TensorFlow and Keras, a Convolutional Neural Networks (CNN) was tested. The goal was to train the model to accurately classify images as either ALL or MM. The model is a simplified version of the VGG-16 model. It was tested with scaling of the original images to dimensions of 20% of the average original dimensions, to 30% of the average original dimensions and then to dimensions of 40% of the originals. Leaky ReLU was used for activation layers. Early stopping was used to avoid overfitting. Because of the considerable class imbalance of MM versus ALL images, class weights were used when the model was fit. The results of testing these various models are presented here and also in this notebook: <https://colab.research.google.com/drive/1uvugMlkCMwXF1j5Len4iCqzCzDKBwkN8> .

The model was trained and tested using a 60/20/20 train/validation/test set split. Model performance was measured in accuracy of classification, although cross-entropy loss was also tracked. The time taken to train each model was also recorded. Additionally, an effort was made to determine which types of test images performed best and worst, that is, which types of images were most and least easily classified.

Consistently, the model runs using 40% and 30% scaling of the images achieved over 95% accuracy on the holdout test set. There is a general improvement in performance and a corresponding increase in processing time with larger size image. It was found that 100% performance was achievable using the 40% scaled images.

