A Peripheral Blood Cell Classification App

Michael Mendez, Juan David Bejarano, Félix González, Julián E Correa H, Camilo Cortés, Santiago Alférez.

Why classify?

- *Blood cell classification is a challenging and time-consuming task.*
- *A safer, more efficient and faster-sorting process.*
- *Automated and reliable process that classifies eight (08) different types of human blood cells.*

Background

Studying cells helps us to understand **how the human body works**.

A peripheral blood smear is essential for the **diagnosis of 80% of haematological diseases**.

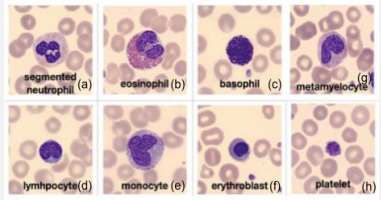
In this project, we will focus on eight (08) specific subtypes of cells who are more frequently observed in **infections and regenerative anaemia**.

Data

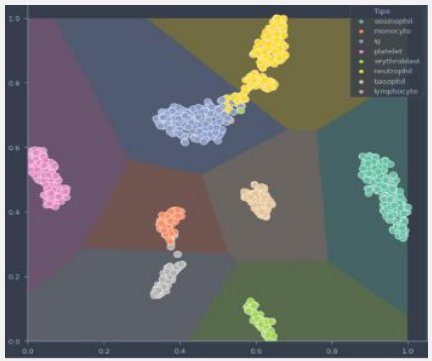
The dataset consists of **17092 digital images** corresponding to eight (08) different peripheral blood cells.

The blood smears were automatically stained using the **May-Grunwald Giemsa technique** which combines the effect of acidic eosin and alkaline methylene blue.

Then the images (**360 x 362 pixels**) were obtained from the analyzer "CellaVision DM96" in the Core Laboratory at the Hospital Clinic of Barcelona.



Images of the eight (08) different types of normal peripheral blood cells.

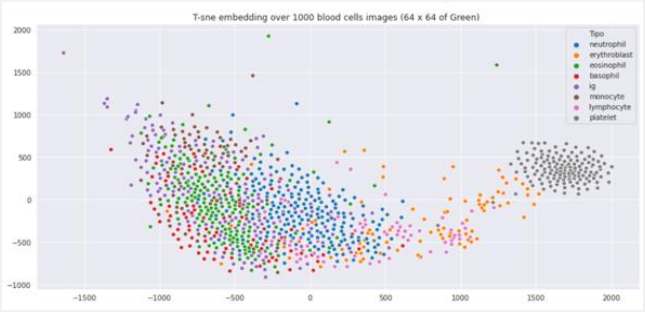


t-SNE over 1290 deep features from the training set.

Exploratory Data Analysis

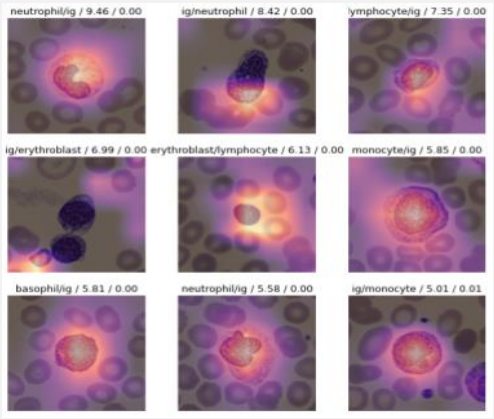
EDA

A useful **method for exploring high-dimensional data** is called t-SNE (*t-distributed Stochastic Neighbor Embedding*). The goal is to take a set of points in a high-dimensional space and find a faithful representation of **those points in a 2D plane**.



t-SNE embedding over one thousand (1,000) cell images.

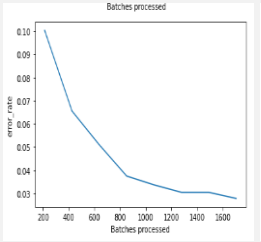
We can see that *platelets and erythroblast* may be different from the rest of the classes. Therefore, based on the previously explained and extra research on the literature, both cell types are the most differentiated respect to the others in size and shape.



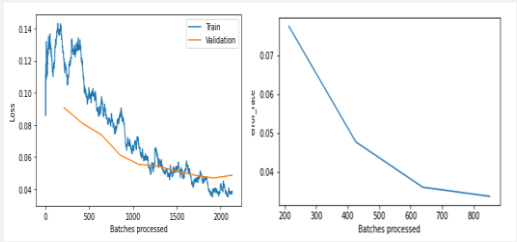
Heatmap of the areas / Grad-CAM of the top Losses which highlights important regions to predict the classes

Model and Statistical Analysis

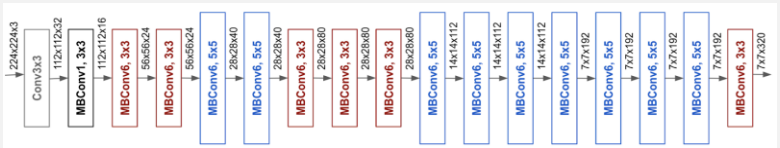
An alternative to training a CNN from scratch is to **fine-tune a CNN** that has been trained using a large labelled dataset from a different application. Therefore, we validated **three different architectures** (*ResNet18, ResNet34 and EfficientNet*) that used pre-trained models and then we made the fine-tuning process to improve the accuracy of the model.



Validation error and performance of the ResNet18 architecture



Results on the ResNet34 architecture without fine-tuning.



EfficientNet Architecture

Results and Conclusions

For each blood cell class, a prediction was generated using the **probability distribution scores** of the CNN's final SoftMax output layer or the tile distribution overlaid on the t-SNE or PCA plot of the CNN's final hidden layer. Afterwards, we applied a clustering technique over the test set and now we have a representation of 512 deep features in 2D.

This work introduces an **interactive and automated peripheral blood cell classification App**. Due to technological advances, the process has become more efficient but it requires great capital expenditure and technical experience personnel. Some health care facilities, especially in rural areas, lack automated analyzer or advanced microscopes. Therefore, a manual and laborious procedure is required which could be prone to errors. Our App offers a solution to this issue.

Using the distribution of prediction scores across the eight (08) classes of our final model, we obtained an incredible accuracy of 99.94% on the test set. Analyzing the top losses, the confusion matrix and the clustering of the deep features, we can conclude that the **neutrophils and the IG cells are the most similar** and more prone to misclassification.

MinTIC. Data Science for All 2. DS4A - Team 14.