University of Technology, Sydney  
Faculty of Engineering and Information Technology

**Classification of Platelets using data from Totally Internal Reflective Fluorescent  
Microscopy**

**by**

**Tegh Bir Singh**

Student Number: 13213929

Major: Data Engineering

Academic Advisor: Qian Su

A 12 Credit Point Project submitted in partial fulfilment of the  
requirement for the Degree of Bachelor of Engineering  
07/11/2022

**Statement of Originality**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **UTS: ENGINEERING & INFORMATION TECHNOLOGY** | | | | | |
| **SUBJECT NUMBER & NAME**  41030 Engineering Capstone  Project | **NAME OF STUDENT(s)** (PRINT CLEARLY)  *Tegh Bir Singh*  *SURNAME FIRST NAME* | | | | **STUDENT ID(s)**  13213929 |
| **STUDENT EMAIL** tegh.b.singh@student.uts.edu.au | | | **STUDENT CONTACT NUMBER** | | |
| **NAME OF TUTOR**  Qian Su | | **TUTORIAL GROUP** | | **DUE DATE**  7/11/2022 | |
| **ASSESSMENT ITEM NUMBER & TITLE**  Final Report | | | | | |
| I confirm that I have read, understood and followed the guidelines for assignment submission and presentation on page 2 of this cover sheet.  I confirm that I have read, understood and followed the advice in the Subject Outline about assessment requirements.  I understand that if this assignment is submitted after the due date it may incur a penalty for lateness unless I have previously had an extension of time approved and have attached the written confirmation of this extension.  **Declaration of originality**: The work contained in this assignment, other than that specifically attributed to another source, is that of the author(s) and has not been previously submitted for assessment. I understand that, should this declaration be found to be false, disciplinary action could be taken and penalties imposed in accordance with University policy and rules. In the statement below, I have indicated the extent to which I have collaborated with others, whom I have named.  **Statement of collaboration**: | | | | | |
|  | | | | | |

**Abstract**

The project title is **“Classification of Platelets using data from Totally Internal Reflective Fluorescent Microscopy (TIRFM)”.** My name is Tegh Bir Singh and for the 2022 Spring semester I have been working on a way classify what stage a platelet is in. This is an important problem as it allows another way to quantify the effectiveness of a drug in terms of platelets and to better understand the behaviour of platelets better with respect drugs.

However, the couple of problems that needed to be sorted before we can actually create an ML model.

1. Image data needs to be collected
2. The platelets need to be segmented before they can be classified
3. As the data is not labelled a way to label the data is needed for supervised machine learning
4. There are issues with classifying the platelets because the overlapping each other.

The data has fortunately already been collected using TIRFM by Qian Su.

To take of (2) we used a python package of cellpose. This creates outlines of the cells by using pre-trained neural network under the hood.

Using the outlines of the platelet we are then able to get the user label what stage each platelet by assigning a number from 1-3 (integers only).

Ultimately, we unable to get the desired results when using support vector machine with low precision and recall of class 3. However, we identified problems with our approach. This offers the opportunity to use the work from this project as a framework for building an ML model.

**Acknowledgements**

I would like thanks my supervisor Dr Qian Su for supporting me throughout the project and for letting work on this project. Without his support and the data, that he has collected this project would not have been possible.

I would also like to thank, Mr Yuekai Sun. Although I have not met you personally the work you have done particularly for segmentation has been interesting to read.

Contents

[1. Introduction 6](#_Toc118739758)

[1.1 Background Research (Project) 6](#_Toc118739759)

[1.2 Background Information (Platelets) 6](#_Toc118739760)

[1.3 Research Question 7](#_Toc118739761)

[2. Literature Review 7](#_Toc118739762)

[2.1 Approaches Considered 7](#_Toc118739763)

[2.1.1 Colocalisation 7](#_Toc118739764)

[2.1.2 AdipoCount 8](#_Toc118739765)

[2.1.3 Other ML Approaches 8](#_Toc118739766)

[2.2 Current Approach 8](#_Toc118739767)

[3. Methodology 9](#_Toc118739768)

[3.1 About the Data 9](#_Toc118739769)

[3.2 Necessary Python packages 9](#_Toc118739770)

[3.2.1 Reading the Data 9](#_Toc118739771)

[3.3 Project Tasks 9](#_Toc118739772)

[3.3.1 Data Visualisation 9](#_Toc118739773)

[3.3.2 Cell Segmentation 11](#_Toc118739774)

[3.3.3 Image Annotation 13](#_Toc118739775)

[3.3.4 Getting Cell Properties 14](#_Toc118739776)

[3.3.5 File Processing Pipeline 15](#_Toc118739777)

[3.3.6 Machine Learning (ML) 15](#_Toc118739778)

[4. Interpretation of Results 16](#_Toc118739779)

[5. Conclusion 17](#_Toc118739780)

[5.1 Summary of Work 17](#_Toc118739781)

[5.2 Potential Future Work 17](#_Toc118739782)

[6. References 18](#_Toc118739783)

# 1. Introduction

## 1.1 Background Research (Project)

Dr Qian Su (aka Peter Su) is responsible for the **Q**uantitative **I**mage **a**t **N**anoscale **Su**per-Resolution

(**Qian Su**) program. Under this program he has developed the “Imaging Profiling Platform for Cardiovascular Disease.”

Due to the extremely small size of platelets super-resolution imaging techniques are needed to see them. In addition, the platelets shape changes extremely rapidly meaning the images need to be captured in a very small-time scale to see the changes over time. As a result, understanding the behaviour of platelets in the treatment of thrombotic disease has been very difficult.

This project can potentially help us to better understand a drug effectiveness through the use of data as it will allow us to see platelets under various conditions and how they are reacting.

As the project is more focused on IT there will be limited interaction with medical aspects in the project.

Hence, we will discuss the research proposal with a focus on IT.

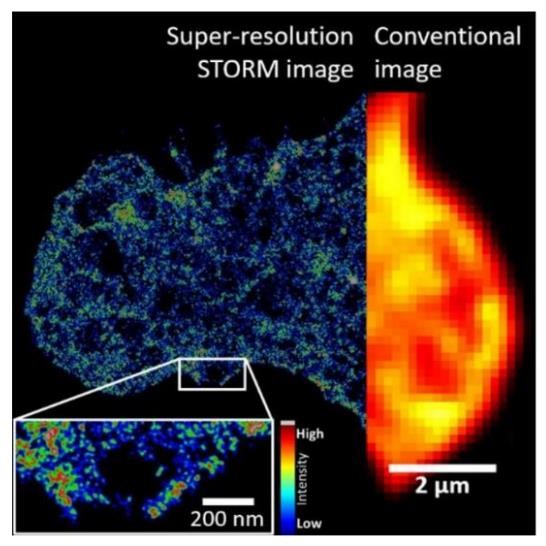


Fig 1: A comparison of super-resolution images on the left and images from a traditional microscope on the right. (Image credit: Dr Qian Peter Su, ‘The Sydney Morning Herald & Sydney Research twitter’). Note 1μm = 1000nm

## 1.2 Background Information (Platelets)

Platelets are fundamental importance in human biology. A small number of platelets can cause excess bleeding while healing from minor injuries. This is because the platelets coagulate together which block the injury. In contrast a high number of platelets can cause thrombosis that is blood clots that can have various complications such as a heart attack.

A person can be said have a low platelet count if for a litre of blood, they have less than 150 x 109 platelets. In severe cases a person will low platelet count with have on order 106 platelets per litre. There are many possible causes that come from illnesses causing such as heart attack, hypotension and dehydration.

Research into platelets can have a profound impact. According to the World Health Organisation (2019), the top 3 leading causes are ischaemic heart disease, stroke and chronic pulmonary disease. If better antiplatelet drugs can be identified the number of deaths caused by these diseases will reduce.

Furthermore, not only can anti-platelet drugs help with cardiovascular disease but also with chemotherapy. Research by Cronkite et. al. (1952) has found exposure to radiation cause cancer. Due to the side effects of chemotherapy anti-platelet drugs are needed to prevent thrombosis.

## 1.3 Research Question

The research topic will be platelet classification. We will be examining how platelets interact with certain drugs that inhibit platelet activation. Once activated the platelet will form a blood clot. This is useful in case of an injury to stop blood flow. However, a large amounts of blood clots can lead to vein thrombosis which in turn can cause various medical conditions, in particular, a heart attack There is no specific research question but there is more of general question. That is:

1. How do platelets react under certain conditions?
2. Of particular interest is how will integrin αIIbβ3, a type of platelet, react with certain drugs such as TGX221 or AZD6482?

A major issue in answering these questions is in how we classify in what stages a platelet is in. This can help us determine whether a drug puts platelets in particular stage and quantify its effectiveness in some way regarding platelet morphology.

So, in order to answer the research question, we will attempt to construct a machine learning model that can correctly classify platelets in different states. In particular

* Inactivated: platelet is at normal size (2μm)
* Semi-activated: Filopodia is released
* Activated: Platelet is enlarged

# 2. Literature Review

## 2.1 Approaches Considered

A past UTS student Mr Sun has already considered some ways to approach the problem in his “Literature Review” for his Engineering Capstone project (2022). We will discuss the issues with some of these approaches.

### 2.1.1 Colocalisation

The ImageJ plugin colocalisation was considered for segmenting the platelets. Unfortunately, due to issues with overlapping and touching cells it was not able to perform well on the data.

### 2.1.2 AdipoCount

Another approach was the use of the AdipoCount algorithm. This algorithm is used for segmenting of adipocytes, i.e., fat cells. Unfortunately, due to the use of the watershed algorithm in AdipoCount the AdipoCount algorithm was not suitable.

### 2.1.3 Other ML Approaches

Although it is not quite the same problem there has been ML model built for segmentation of “Deep Vein

Thrombosis” (DVT) by Huang et. al. (2019). This used artificial intelligence, in particular

CNN (Convolutional Neural Network) model in conjunction with a encoder-decoder network. The CNN is used to create the boundary box for the segmentation while the encoder-decoder network is used for processing high dimensional data and converting it lower dimension. This makes it easier to develop the machine learning model. Without the auto-encoder there will be more parameters in the data to work with which in turns increases the chance of the model overfitting.

So, it is possible to employ a similar solution. It should be noted as we are concerned with individual platelets it will harder to perform image segmentation on integrin compared to thrombi due to the nanometre size of platelets.

An interesting solution was used by Clauser J. et. al. (2020). They used a machine learning algorithm called random forest to classify blood types from platelets. Fluorescence microscopy was used to create the images, which were classified overall with 93% accuracy.

This is quite surprising as they are using classical machine learning techniques instead of neural networks. Generally, classical machine learning algorithms will not perform as well as when there are large amounts of data. This could perhaps mean that the approach using random forest for image classification could be successful.

In order to deal semi-supervised streaming data Zhang Z. et. al (2021) developed a machine learning algorithm to deal with “real-time noisy, sparse, and ambiguous images of platelet”. This work was done in other label platelets and see structural changes, which happen rapidly. This has had great impact as other earlier approaches were not able to accurately identify structural changes in the platelet due to technical constraints.

So, it may be possible to potentially use the work of Zhang Z et. al in the construction of the machine learning model but in order to identify platelet activation.

## 2.2 Current Approach

The approach that was undertaken by Mr Sun was use a website called cellpose. There is also a Python package called cellpose which can be used instead. The website allowed image analysis to be done online without any Python knowledge however the image need to be uploaded manually.

**Issues with Current Approach**

Since a web GUI is being used for the data analysis in cellpose this means that it is not practical to analysed thousands of images. As a result, Python code is needed to automate the image analysis so this task can be performed.

# 3. Methodology

## 3.1 About the Data

All the data has already been collected by Dr Qian Su.

The data set is roughly 10GB and each file is divided in 10 files so each file is approximately 1GB. The data is stored on Google Dropbox which can be downloaded through GUI (Graphical User Interface).

* This means no data engineering skills will be required to download or uploading the data to a database.
* Due to the large size of the data, it could be difficult to work with all the data at once. For our purposes this will not be a concern as we are working on a subset of the data.

For each image there are three channels, SZ22, MBC370, PAC1. Each of these represent a different fluorescent staining which is used to capture the images. These channels represent the different states of a blood cell in the data, whether they are resting, semi-active or active respectively.

Each file has 300 frames. Each frame represents a single point of time where an image has been taken. The file extension is “tif” also known as “tiff” which stands for “Tag Image File Format”.

**Important Note: All code will be given to Dr Qian Su.**

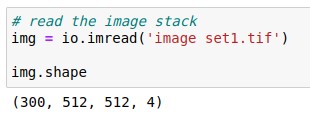
## 3.2 Necessary Python packages for Reading the Data

* matplotlib.pyplot
* numpy
* skimage.io

### 3.2.1 Reading the Data

Using the io module from the skimage package we find the file contains 300 images each of which contain 512 by 512 pixels. The img object is numpy array. The 4th dimension of array contains the 4 different fluorescent channels of the image.

Important note: Only the 1, 2, 3 index of the 4th dimension of the array are useful for the data visualization. The 0th index contains miscellaneous data.

**Figure 3.2.1.1: How to read tif file using skimage.io**

## 3.3 Project Tasks

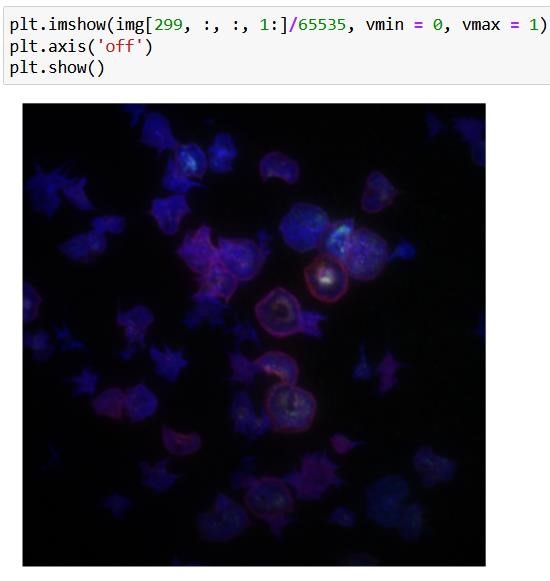
### 3.3.1 Data Visualisation

Using matplotlib.pyplot we can plot the image. Since the data stored in the tif file is 16 bit we set vmin = 0 and vmax = 65535 (i.e., 216 - 1) to control the brightness of the image.



**Figure 3.3.1.1: Image of the first fluorescent channel, frame 300**

Out of interest we have also plotted an RGB image with the three channels combine into one image.



**Figure: 3.3.1.2: RGB image, of platelets. Please note colours are artificially added by Python**

It is interesting to observe the reddish illumination of some of the cell’s edge.

To further improve the visualisation, we have created a GIF. This allows see what happens to platelets over time.

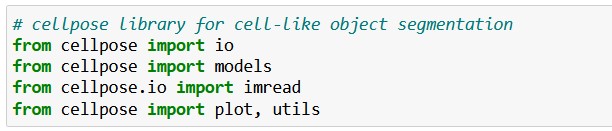
Text

Description automatically generated

**Figure 3.3.1.3: Code to create a GIF from an image**

### 3.3.2 Cell Segmentation

For the image segmentation we used cellpose a python library to perform image segmentation (cellpose developer, Stringer. C and Pachitariu. M, 2020). cellpose is used to perform image segmentation of cell-like objects. We import the following functions from cellpose to perform the segmentation.



**Figure 3.3.2.1: Function from cellpose needed for segmentation**

A big issue is cellpose can only handle grayscale images however in the data we got images in 3 dimensions. As it turns out the only important channel is 3rd index. Another alternative is using a function that converts RGB images to grayscale.

We have taken both approaches. As it turns out did not have make much impact on the segmentation results due to the predominance of 3rd fluorescent channel.

To begin the segmentation, we need to select the model type. As platelet do not contain nucleus we select “cyto” for cytoplasm.



We then read the file. Due to the way the imread function is defined in cellpose we need to index differently to select the last frame and 3rd fluorescent channel. This is slightly different to the way it is done in skimage.io.



**Figure 3.3.2.2: Code to process the tiff file in the correct format for segmentation**

We now create the various variables needed for segmentation file that will be used for segmentation.

The channels keyword argument is set to [[0, 0]] to indicate the image is be treated as grayscale. We set diameter to 40. We found this value to be the most suitable after experimentation.



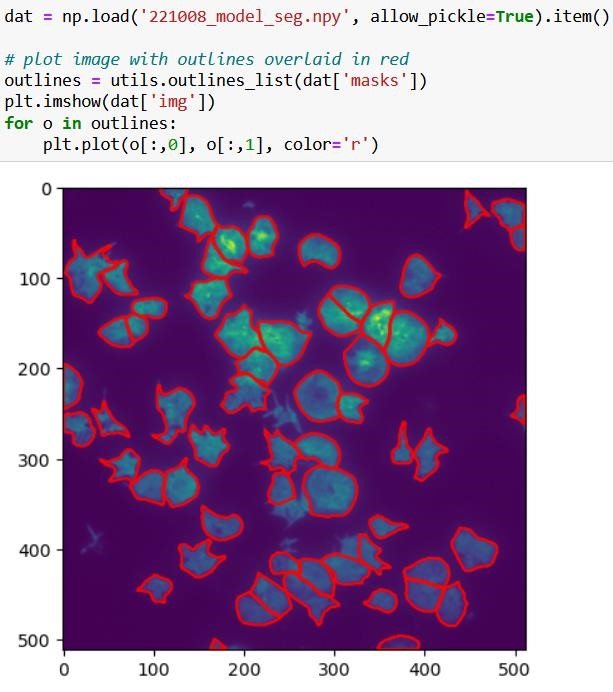
We then create the file that will be used to segment the image.



We then use the following code to plot the boundaries of each cell. We can see clearly that some cells are clearly being missed however most cells are segmented properly.

By examining the image more closely we can see that:

* There is some mis-segmentation due to some cells been too big. For example, take a look at (150, 300).
* Some of the cells are overlapping. See cells around (150, 350). This will cause us troubles in developing an ML model to classify platelets



**Fig 3.3.2.3: Image of cell segmentation using cellpose in Python**

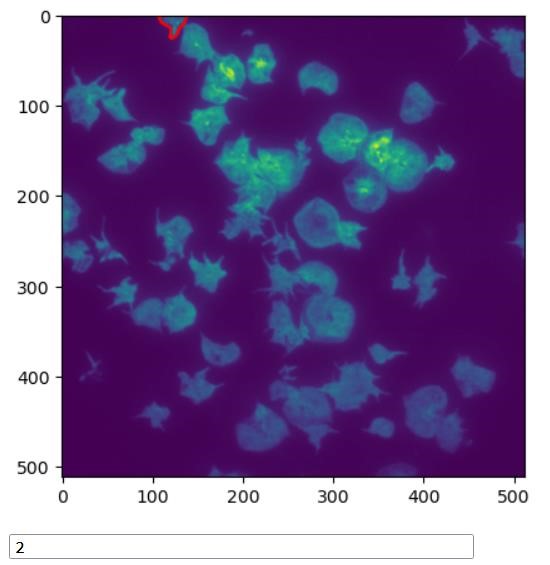
### 3.3.3 Image Annotation

With the images now segmented we now need to annotate the images. The code uses the segmentation file that we have created earlier. We then plot one of the outlines of the cell.



**Fig 3.3.3.1: Code use to annotate images**

So, we will see something like image below. We can then input a number, from 1-3. 1 being resting state, 2 is filopodia is clearly visible, and 3 is cell fully enlarged Using this we can for each cell assign a classification.

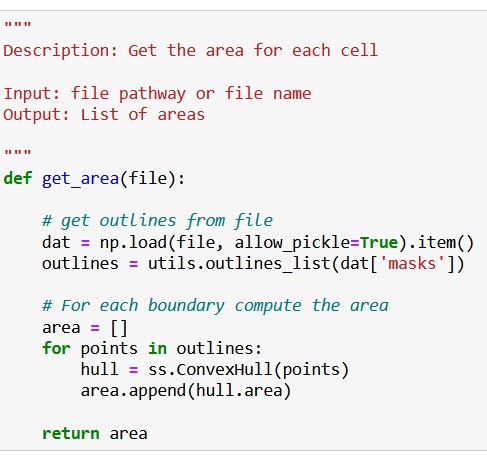


**Fig 3.3.3.2: Function for annotating image being executed.**

### 3.3.4 Getting Cell Properties

Now we need to get the cell properties so we can perform machine learning.

We import scipy.spatial as ss. We will use this find the area of platelets.

Using the following code, we can get the area for each

**Figure 3.3.4.1: Function get\_area**

### 3.3.5 File Processing Pipeline

To automate the processing of the files we have created a file processing pipeline that can process the files in one go.

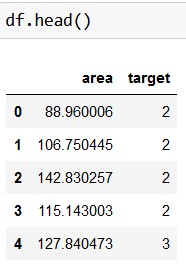
To do this we use a function called get\_all\_filepaths. As the code is not important, we have left it out



Using this we can create function called file\_processing\_pipeline. This returns a dataframe



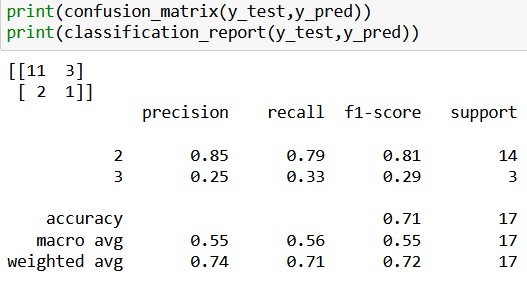
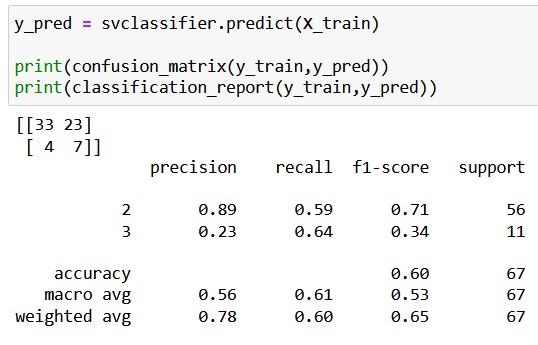
The dataframe looks as follows:



**Figure 3.3.5.1: Output of dataframe after using file\_processing\_pipeline**

### 3.3.6 Machine Learning (ML)

The data is imbalance. 1/6 of the data is in class 3 and the remaining is in class 2. This can cause issue with ML model as it can mean they predict all the cells in class 2 and get accuracy of approximately 83%. To deal with this we used a weighted SVM, so misclassifications in class 3 are penalized more heavily. The results for training and testing data are shown respectively as follows:



**Figure 3.3.6.1: ML results for classification of platelet data**

# 4. Interpretation of Results

The results are not that good as we can see from the precision and recall of class 3.

A couple of potential things could be going on:

* The data is not being labelled well. We need establish better which cells are in which class. It could be a cell is in mix of both class 2 and 3
* Overlapping area is causing issue in the classification of cells. This means the area assigned to each cell is not correct
* A more thorough analysis of the segmentation is needed to ensure that the cells are correctly being segmented

As we can see more work needs to be done.

.

# 5. Conclusion

## 5.1 Summary of Work

In the process of trying to class platelet into different stages we needed to:

* visualise the data
* segment the data
* annotate the data

These processes enabled us to develop a ML algorithm to perform classification. Unfortunately, we were unable to get the results we were after as the precision and recall were quite low particularly in the testing data.

However, the work we have discussed have done, can be used as a framework to improve the results that we have obtained. In particular, the mechanism we have developed for labelling the data can be used so that a supervised ML algorithm can be used.

## 5.2 Potential Future Work

The most major improvement to be made is in improving the results of the ML. To improve the results, it will be to relabel the images, e.g., in 3.3.2.3. The reason for this is the red outlines used to show the platelet in the image make it hard to see a clear outline of the platelet being labelled.

So have one image with the outline and the actual image without the outline shown side by side.

Another potential improvement will be labelling cells on the border as NA.

Figuring out a way to also properly segment cell will enable further improvement. Particularly in case where a large cell is segmented as two cells. Perhaps a better way to approach this would be manually fix the cells that are not correctly segmented.

It is also possible the ML results desired may not be enough and tools like 3D microscopy may need to be used.

# 6. References

Clauser, J. C., Maas, J., Arens, J., Schmitz-Rode, T., Steinseifer, U., & Berkels, B. (2020, November 6).

*Automation of hemocompatibility analysis using image segmentation and supervised classification*. Engineering Applications of Artificial Intelligence. Retrieved April 8, 2022, from <https://www.sciencedirect.com/science/article/pii/S0952197620302980>

Cronkite, E. P., Jacobs, G. J., Brecher, G., & Dillard, G. (1952). The hemorrhagic phase of the acute radiation syndrome due to exposure of the whole body to penetrating ionizing radiation. The American Journal of

Roentgenology, Radium Therapy, and Nuclear Medicine, 67(5), 796-804.

Huang, C., Tian, J., Yuan, C., & Zeng, P. (2019, June 9). *Fully automated segmentation of lower extremity deep vein thrombosis using convolutional neural network*. BioMed research international. Retrieved April 8, 2022, from <https://pubmed.ncbi.nlm.nih.gov/31281832/>

Hunt, B. J. (2008, March 1). A*wareness and politics of venous thromboembolism in the United Kingdom*. Arteriosclerosis, Thrombosis, and Vascular Biology. Retrieved April 8, 2022, from <https://www.ahajournals.org/doi/10.1161/ATVBAHA.108.162586>

Sun. Y. *Imaging Profiling Platform for Thrombotic Disease and Anti-Platelet Therapeutic.* Date: 2022, Feb, 18

Stringer. C, Pachitariu. M . *cellpose*. (2020) <https://cellpose.readthedocs.io/en/latest/outputs.html>

World Health Organisation (2019). *The top 10 causes of death.* Retrieved May 22, 2022, from <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>

Zhang, Z., Zhang, P., Wang, P., & Sheriff , J. (2021, March 11). *Rapid analysis of streaming platelet images by semi-unsupervised learning*. Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society. Retrieved April 8, 2022, from

<https://pubmed.ncbi.nlm.nih.gov/33798915/>