**Individual Portfolio**

**Colin Chan**

**ADS2002: Data Challenges 4**

**Project: Medical Image Multi-Label Classification**

**INTRODUCTION:**

This is the individual portfolio of my personal contributions made towards the group project. It will contain weekly entries that outline the tasks that were performed throughout the week, an update of the current state of the code and particular research or insights of the project. It will also contain two separate reflective entries, one halfway through the semester and one at the end of the semester that will summarise the two halves of the semester. The weekly entries will begin from week 2 and end in week 12.

**WEEKLY ENTRIES**

**Week 2**

The catheter placement project was an interesting project that I was very enthusiastic about as it contained analysis and deep learning techniques that were not taught in previous ADS units. Although it would be challenging to use these techniques and with the added difficulty of having a large dataset, I was still excited to test my limits on how much I can deep dive into this image analysis project.

We started off meeting with our project mentors. Simon was particularly helpful with explaining the dataset, how we could approach the project and providing guidance on how we can model the data. I took notes on ideas as they were being discussed by Simon. It was a little bit overwhelming to hear new deep learning jargon, but it was clear what we could and should do.

This week no code was developed as we awaited the dataset to be released by Simon. The team discussed possible approaches to the project and what needed to be researched. We agreed that an orthodox approach should still be taken; that is to clean the data, perform exploratory data analysis, fit a model, and test the model. The team has also agreed to use discord as the main source of communication, but occasionally check-in on teams to ask Simon questions or look out for announcements.

**Week 3**

This week the team’s Github repository was established by Huda in class and the team could begin to view the dataset on Google Collab. As this was the first time we were dealing with images and it was unfamiliar territory so I looked at some notebooks on exploratory data analysis performed by other people on Kaggle to get a better understanding of the EDA process with images. This was reported back to the team, and we were able to do similar EDA on the data.

While looking at the data, we noticed that there were 2 features that did not align with what the team decided on trying to predict. The “NGT – Incompletely Imaged” and “Swan Ganz Catheter Present” columns. From further reading, the incompletely imaged are images that did not capture all the necessary details or information and the Swan-Ganz catheter was a specific type of catheter for the right side of the heart. The model we are trying to achieve is to be able to distinguish whether the catheter has been inserted incorrectly. This means that incomplete images are not required and we were trying to accommodate for all types of catheters not just the Swan-Ganz catheter. These were dropped from our data frame which resulted in the following state of the data frame:

A screenshot of a computer

Description automatically generated

From the EDA of the data, we were able to generate a bar plot of the distribution of catheter types and positions. This gave us a really good idea of the distribution of the data.

A screenshot of a computer

Description automatically generated

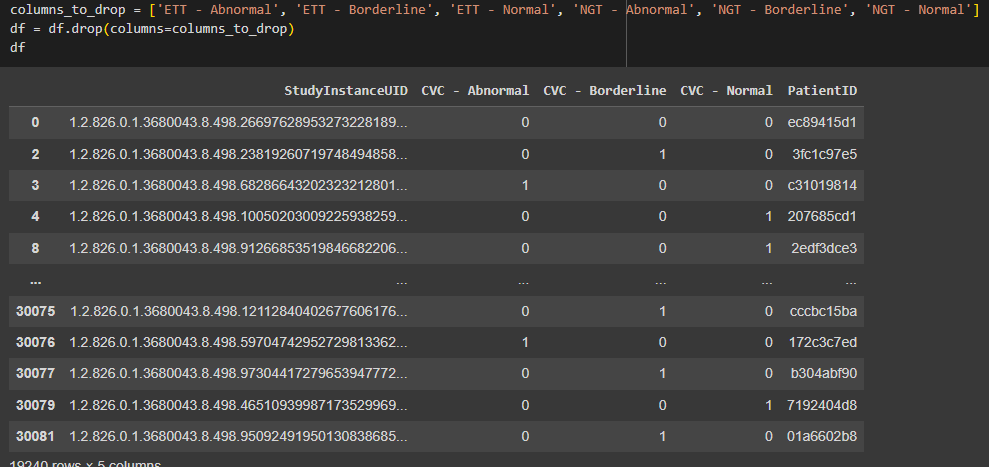
This meant that for our model to be accurate, we need to balance this data to avoid model bias. Balancing techniques were set as a task for all members in the team to research and report next week.

**Week 4**

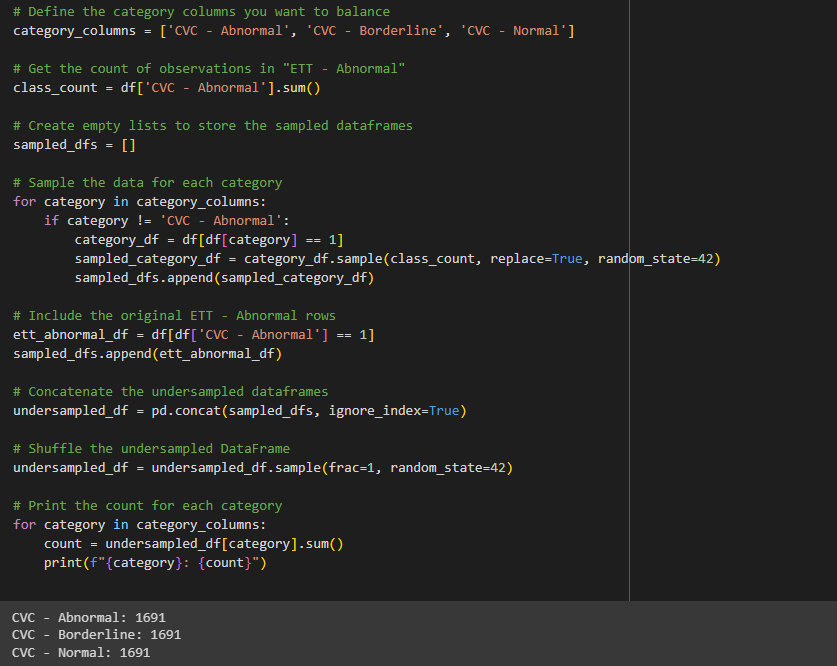
From the exploratory data analysis performed last week, we realised that there was an imbalance in the dataset. This week’s focus was on finding methods to balance the dataset so that we can use the data for training and testing. As a team, we discussed a few ways we can balance the dataset. Huda and I completed a virtual internship for British Airways and learnt how to balance the dataset by over and under-sampling. Zach also suggested to the team SMOTE which is an oversampling technique for imbalanced classifications.

When attempting to perform under-sampling balancing, I realised that it was different to the task we had in the virtual internship as there were multiple types of catheters and in different locations, they all had varying numbers of observations and there was no comfortable number to under-sample all the categories to. It was then decided by the team that we would simplify our model to only the CVC type catheter as it had sufficient samples and then progress to more complex models later.

So all the other columns were dropped and we looked into the SMOTE technique and attempted to balance the data by under-sampling.



**Week 5**

This was the code that was committed to the repository for balancing the dataset via under-sampling.

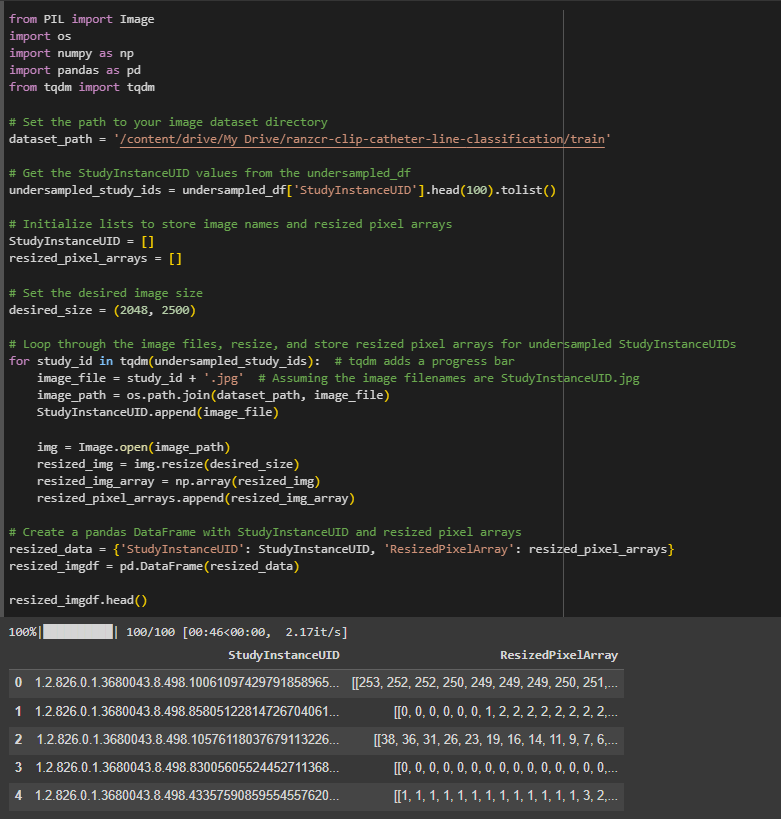
The dataset is now balanced with 1691 samples in each position of the CVC catheter. The next stage was to implement a simple model with the current dataset that we have. We faced an issue with storing the image names and the resized pixels arrays into the dataframe, as it took 45 minutes to run and loop through all the images to convert them. Huda managed to run this and afterwards attempted to save the file as a CSV to send to the group to use that dataset for the models, but it had /n on all the line breaks in the array and it was not usable even after we tried to run a loop to remove the “/n” or tried to replace them with “ “.

This week we also went through feature selection, but as a team we already removed some features from the very beginning and did not see this topic to have much relevance to our project.

We were not able to obtain a dataset to train models on during class and I had scheduled a team meeting for Thursday night to see if we can resolve it before next class.

**Week 6**

We were not able to solve the issue with our dataset during the team meeting, but in class this week we worked with the other catheter group project team to see how they attempted to obtain a dataset to train models on. We were able to get some help from the other team and were able to convert the image names and resized pixel arrays into our dataframe. The below is the code committed:



This block of code was to merge the dataframes together which generated merged\_df for our team to train models on:

A screenshot of a computer program

Description automatically generated

The next step was discussing what models we would try to use. A few ideas were tossed around and I was allocated the task of learning and researching neural networks and using tensorflow to train a model.

**Mid-Semester Reflection**

It has been a challenging yet fun journey so far working with the team on this project. Previously, my experience with working with other team members is to divide up all the tasks and allocate tasks to different people and once everyone has done their part, we come together to combine everything together. However, not all complex projects can be completed or approached the same. I have learnt that with my recent team projects, there are some stages of the project that needed to be completed together as a team, so the team has a better understanding and alignment of what is happening. For example, we were not able to research and learn what models were suitable for our data, until we have a clear understanding of what data we are working with, and this was only available once the data has been cleaned and explored.

We have encountered a few problems in the project so far. One of them being the issue with getting a complete dataset to begin training models. Not only is the dataset the largest dataset we have ever worked with, but it also contained elements such as image classification that we have never learnt or been exposed to. This played a big impact on our performance. I managed to overcome this by encouraging the team to be more open when asking for help, even if it meant asking the other group that is doing the same project. It was quite difficult to do this as there was always an underlying feeling of competitiveness between teams who are doing the same project.

Areas I can work on are definitely the communication. As Huda, Simon and I have worked together previously and are friends outside of class, we tend to have small talks of the project outside of class and only mentioning major details or points to the entire team. While it is difficult, it would be beneficial to organise more team meetings where possible to align the team more together and keep each other accountable for our work.