**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

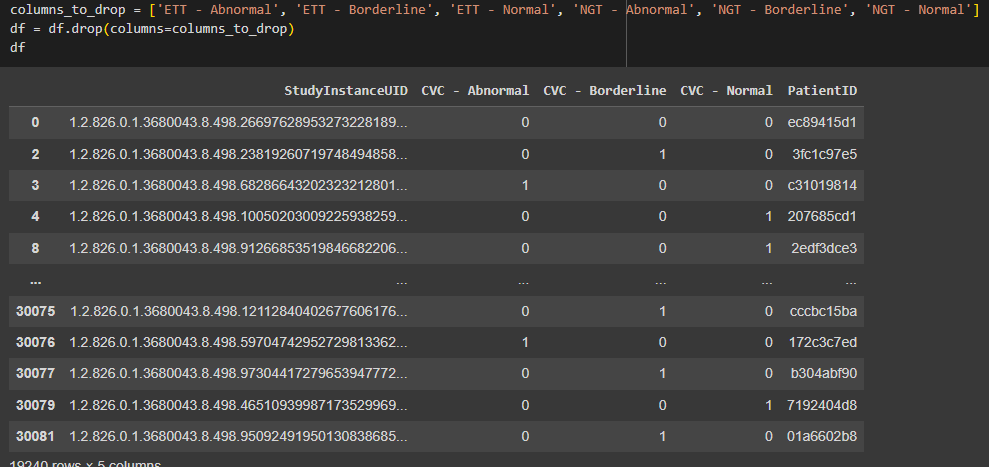
This meant that for our model to be accurate, we need to balance this data to avoid model bias. Balancing techniques were set as a task for all members in the team to research and report next week.

**Week 4**

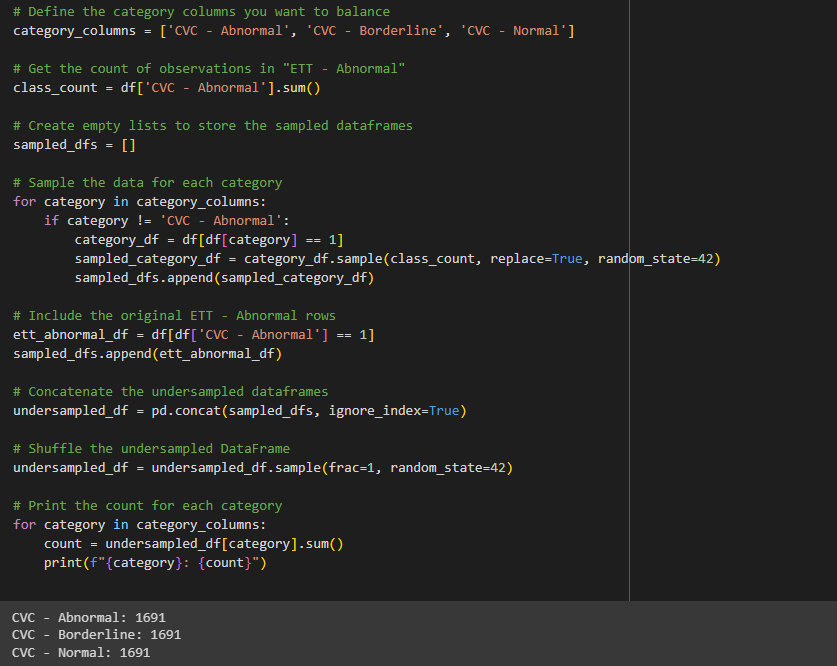
From the exploratory data analysis performed last week, we realised that there was an imbalance in the dataset. This week’s focus was on finding methods to balance the dataset so that we can use the data for training and testing. As a team, we discussed a few ways we can balance the dataset. Huda and 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.

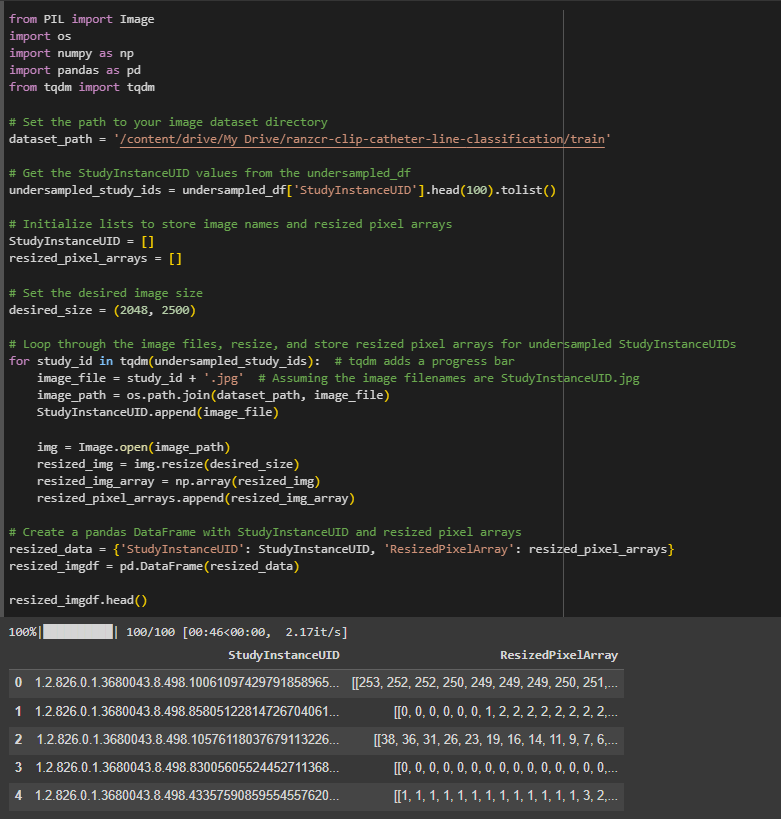
The dataset is now balanced with 1691 samples in each position of the CVC catheter. The next stage was to implement a simple model with the current dataset that we have. We faced an issue with storing the image names and the resized pixels arrays into the dataframe, as it took 45 minutes to run and loop 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.

**Week 7**

Despite learning about perceptrons, the building blocks of a neural network, it was difficult to grasp how we could incorporate this type of modelling into the project. This week was spent trying to learn and figure out how to delve deeper into neural networks and how it can be used.

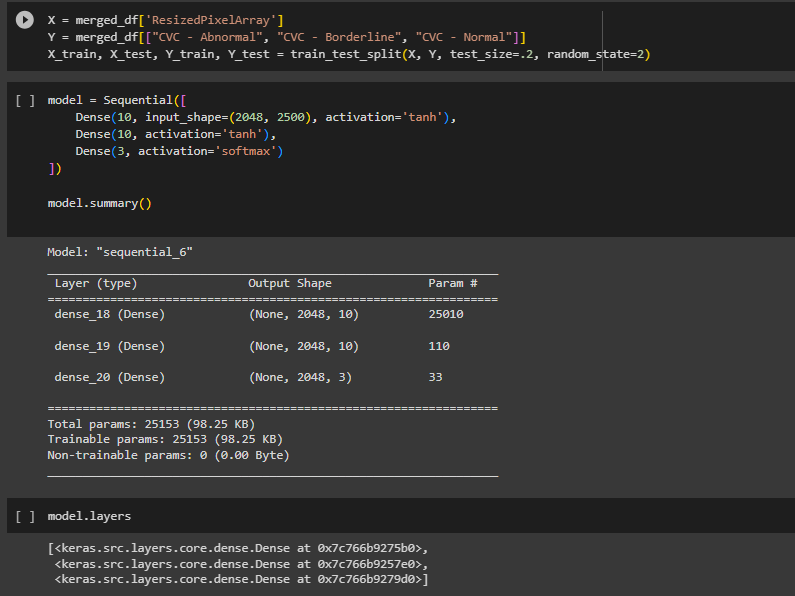
Through some research, I have found that neural networks consists of nodes that is organised by layers, starting with an input layer, then one or many hidden layers and finishing with an output layer. The training of a neural network is simply just the adjustment of weights of each predictor or feature in the data. However, because our dataset is mainly images, we need to be using a convolutional neural network, which is just a specialised type of neural network to process data like images. The difference between a neural network and a convoluted neural network is the types of layers it contains.

With this information in mind, I was able to follow some of Simon’s suggestions to look into tensorflow, ResNet, DenseNet and UNet. There were a lot of websites that explained the concept of neural networks, but it was still a little difficult to fully understand.

The mock presentation is next week and I helped with organising everyone’s presentation, to see who was going to talk about what part to ensure that our presentation flowed smoothly along with the slides.

**Week 8**

This week, with some help from Huda and Simon, we managed to create a very basic neural network with only one hidden layer on only the CVC catheters. This was because we only used the under-sampling to balance the dataset and did not give a genuine attempt to over-sample with SMOTE. Below was the code committed:



A screenshot of a computer

Description automatically generated

This was a decent start because neural networks were a complex concept to understand.

After this mock presentation, the group received feedback regarding the research questions of the project. It was not clearly defined, and the relevance was questionable, so I took this opportunity to explain my approach of defining the research question. The group was able to come up with better research questions and the presentation slides and report was adjusted.

**Week 9**

My application for MASSIVE was approved this week and the majority of this week was spent transferring the Google Collab notebooks across to MASSIVE. However, there were some issues with loading the dataset and libraries into the Jupter Notebook on MASSIVE. After sitting down with Simon, we managed to figure out how to fix the working directory to property reference the dataset source file and to set the notebooks working directory to it. We also figured out how to properly pip install the libraries required for our project. All this knowledge was passed onto the group and I helped Huda, Simon and Joanne to properly configure their MASSIVE for the project notebook.

This week I also started researching into ResNet-50 and looked into the documentation found online. Huda managed to produce a ResNet-50 mode and the following code was committed:

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

**Week 10**

This week the group finalised the allocation of presentation parts and I was allocated the exploratory data analysis. Joanne created the shared google document for the report and I created a skeleton template for what needed to be part of the report based on the report guidelines.

I began analysing the EDA we had found and also updated the graphs so the visuals would look better for the presentation. The below code was committed:

A screen shot of a computer program

Description automatically generated

A black and white background with white lines

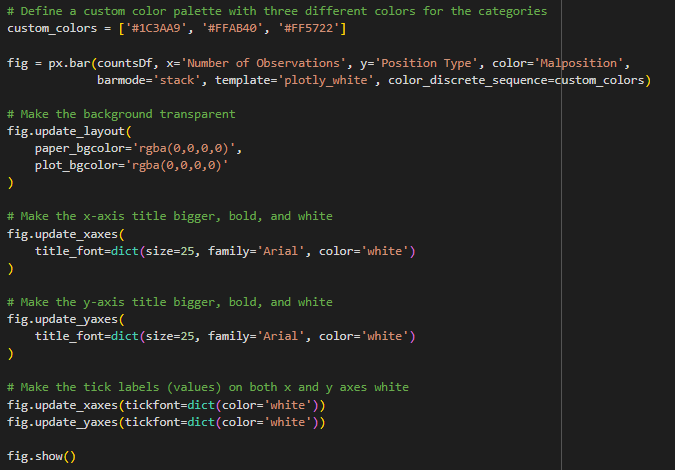
Description automatically generated

This graph was the distribution of the number of observations per patient ID. There were not many relationships between the features, but I found that this relationship would be a noteworthy one because repeated sampling of the same patient would mean the data would be bias and would introduce underfitting to the model because there is not enough underlying patterns for the model to capture and learn. Fortunately in our case, the distribution was very promising because it was positively skewed and the number of observations per patient ID dropped significantly after 10 observations so we did not have to specifically do anything to mitigate the bias.

**Week 11**

This week was finalising the presentation slides and reporting. I started to organise team meetings on Discord to ensure that everyone knew what everyone knew what they were doing for the presentation and report, making sure that the team was align with our approach to the presentation etc.

I continued where I left off last week and finished my research on balanced and imbalanced datasets. This was the code that was committed:



A black and orange graph

Description automatically generated with medium confidence

Although it was obvious from the start that the dataset needed to be balanced, I managed to finish the EDA section of the reporting and prepared for the presentation for this section of the EDA. Through my research I found the consequences of using an imbalanced dataset and how it can cause the model to skew towards the majority class (CVC catheters) which would then cause overfitting to the model because the model would learn all the patterns and relationships in the CVC images, and when it comes to predicting other catheters it would perform poorly because it was not fed enough information for those types of catheters.

**End of Semester Reflection**

This project has definitely been a wild rollercoaster ride. There were times where the content was very interesting and motivating to learn neural networks ahead of class, and there were times where it was a constant struggle between too complicated and time-consuming to learn complex neural networks for the project to get better model evaluation scores.

In terms of what I have learnt from working with a team, I have learnt that it is difficult to coordinate with multiple team members. Every team member has different work commitments and some team members had different motivations towards the subject which led to lesser contributions from them. I have also learnt that each team member had different strengths and weaknesses and the project work was more productive when team members were working towards their strengths. This was very interesting to witness, because some team members were good at coding and managed to grasp the logic and new concept very easily, whereas some team members were better report writing or creating presentation slides.

Some of the challenges of the project was trying to learn a new skill or technique that had a very high learning ceiling. Learning neural network modelling can be fun but also challenging. The initial learning of neural networks was very overwhelming, but slowly as the content was taught in class and from more exposure to neural networks, it started to become easier to understand. While there are still lots to learn, the basic fundamentals were less overwhelming.

I had an all-rounder contribution to the project and participated in a little bit of everything, from coding and report writing to organising in-person meetings. I also feel like I played a leadership role in the group without the title or status of a leader, not only from organising team meetings and aligning the team with updates of the projects, but it was from my engagement with each team member, checking in on them to see if they needed any help with what they were working on and prompting discussions with them to bounce ideas and give second opinions. I would like to say that this helped the group run a lot more smoothly.

Upon reflection, I feel like my area of improvement would have been to have more of a focus on coding. As much as other aspects of the project were important, being an all-rounder for the project felt like I was as the saying goes “jack of all trades, but master of none” and I did not learn as much as I wanted to for the coding aspect for this project. This is especially because I selected this medical imaging project to be more exposed and have a closer look at modelling with images. For future projects, I would go into the project with more specific goals and think about what I would like to accomplish upon completion of the project.