



## **School of Science**

COSC2673 – Machine Learning

### **Assignment 2**

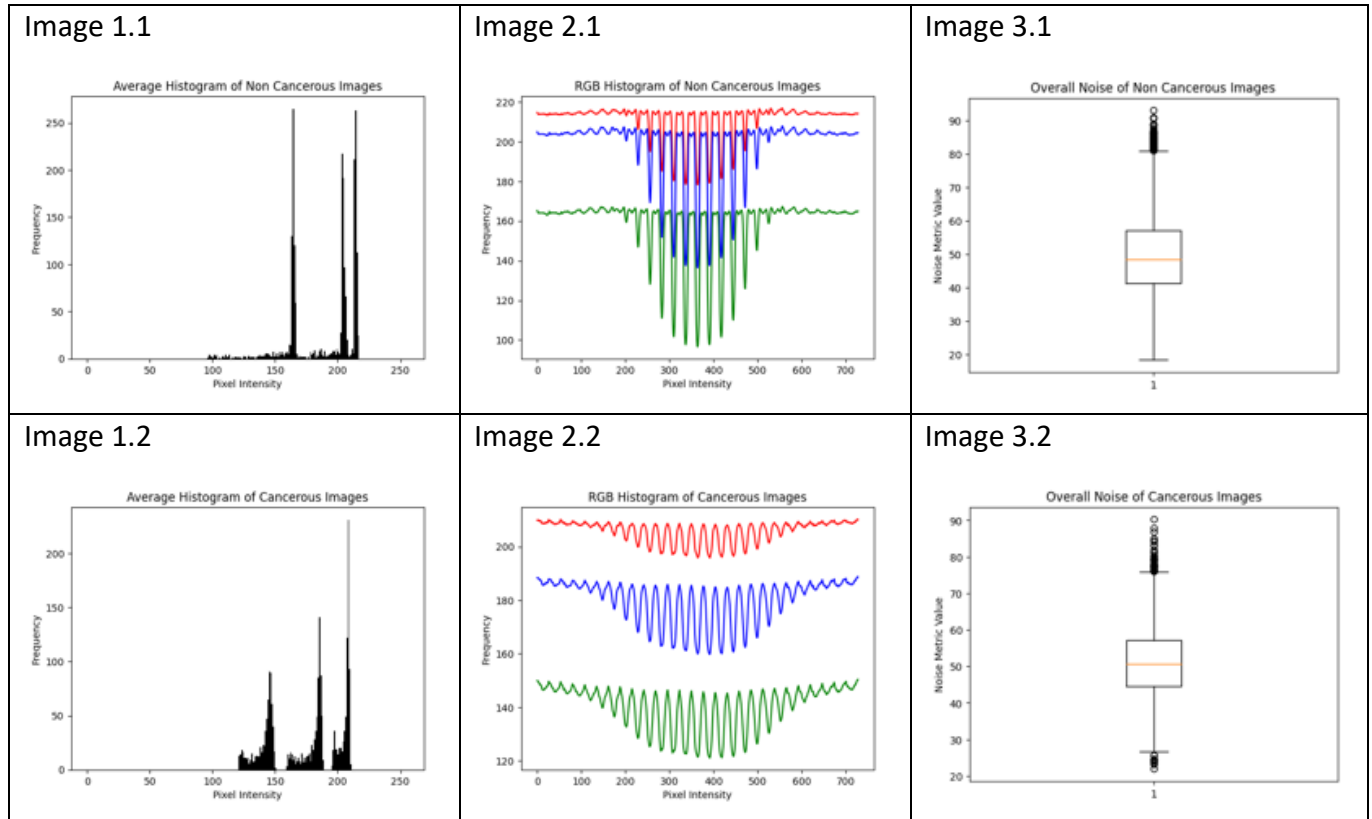
Lecturer:	Dr. Azadeh Alavi
Students:	Kashyap Chilivoju S3659816 Andrew Scicluna s3601842
Submission Due Date:	16/05/2023

## Introduction

Assignment 2 involved developing a machine learning system to classify histopathology images of cancerous cells.

## Investigation & Findings

### EDA



Some key findings within initial EDA were that:

- mainData.csv only contains epithelial cells which are only cancerous and vice versa, meaning no non-epithelial cancerous cells data points are given in the dataset, limiting the data pool, and only restricts it to one cellType, potentially skewing the results of our model, so we must choose our model with this in mind.
- Another finding was that there are multiple instances of the same patient, some patients had a high frequency of images, and some with low. This can cause data leakage while splitting data into test and train set, and so this needs to be acknowledged and worked around.
- There was a lot of noise in the images, more so in non-cancerous images, this can influence the model performance if this is not considered.
- There were twice as many non-cancerous images as compared to cancerous images, while cellType images vary with ratio 2.85: 1.7: 1.3: 1, majority being epithelial cells, and minority being others.

### Pre-Processing

Initial Data preprocessing was performed on the data was a custom data split. This can be seen in appendix 2. The data was split based on Patient ID to ensure no data leakage, using the inbuilt Keras data splitter could mean that a patient's images would be seen in training, validation, and testing datasets.

Due to noise within images as per images 1.1 and 2.1, we implement Data Augmentation in Deep Learning Model 3 to improve on previous models, so to remove as much noise as we could from the cell Images.

## **Approach**

We decided to implement Neural Network as the Baseline Model for several reasons, namely:

- Complexity, non-linearity, and unstructured nature of cell images, and how Neural Network provides a means of capturing these complexities and mapping it out in a neural network is appropriate for problem at hand.
- The ability to pick out patterns, generalize, and map it on a network to classify.
- By using this Pattern Recognition to our advantage and overcoming the issue of restricted dataset of only epithelial cells being cancerous, as neural network's generalization helps find the pattern of isCancerous cell rather than an epithelial cell and correlating it to isCancerous.
- Ease of implementation, as setting up a Neural network is quick and relatively easy to get set up and running, with minimal human input, especially for a baseline model.
- Delivering an output in a reasonable time with fast process times, and reasonable system requirements.

From here, we moved on to Deep Learning Algorithm for the following reasons:

- Deep Learning eliminates the need of pre-processing, and data skewedness.
- Considering both the students doing this assignment lack an understanding of subject matter i.e., Cell Images, Types, Cancerous properties, etc., Deep Learning can work despite having no knowledge on these topics.
- Having a reasonably big number of hidden layers means all the intricacies of each cell can be recorded and have an impact separately within the prediction, in other words, Deep Learning accommodates and acknowledges all the tiny intricacies and distinct features into hidden layer neurons and can make a much-experienced prediction with all the small details in mind.

## **Evaluation of Performance**

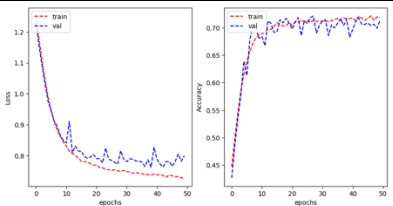
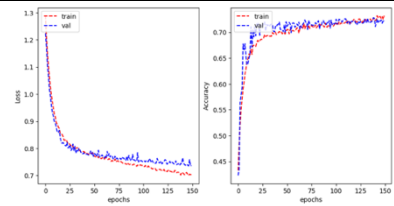
After having implemented a baseline Neural Network, the accuracy came out to be 0.5970 for cellType Classification, and 0.6714 for isCancerous Classification. Furthermore, the loss was 0.9669 and 0.6802 respectively. From here, we must now come up with further optimization, and potentially look at other algorithms to further improve model performance and accuracy.

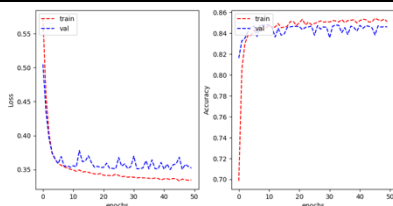
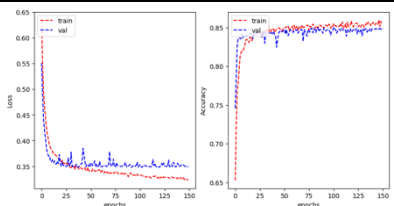
We noticed there was an overfitting with dataset. To combat this, we used L2 regularization and dropout to explore optimization possibilities. These can be seen image above. Throughout our analysis we used learning curves within Keras to plot the output of our models. These are all listed in appendix 4 below. By plotting the training and validation error as a function of the training set size, a sense of the model's bias, and variance can be determined.

Looking at these graphs and at the result of the Test – Loss and Test – Accuracy results we were able to determine drop out to be the best performing model for our use case.

Our model was able to be improved in both classifications using Dropout with a reg\_lambda value of 0.01 for Cell Type Classification and 0.001 for Cancerous Classification.

## Ultimate Judgement & Analysis

Cell Type Classification	
Baseline	Dropout
	
Base - Test Loss: 0.9669 Base - Test Accuracy: 0.5970	Dropout - Test Loss: 0.9035 Dropout - Test Accuracy: 0.6342

Cancerous Classification	
Baseline	Dropout
	
Base - Test Loss: 0.6802 Base - Test Accuracy: 0.6714	Drop - Test Loss: 0.3754 Drop - Test Accuracy: 0.8268

Exploring relevant literature that utilised the same ‘CRCHistoPhenotypes’ Data set, it was evident that in most cases, some versions of a Convolutional Neural Network were utilised. Sirinukunwattana Et al. used Spatially Constrained Convolution Neural Network (SC-CNN) on the same image dataset “Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images” [1]. And for the classification of a nuclei, it was proposed to use a Novel Ensemble Predictor (NEP) coupled with CNN. They found that this approach produced the best F1 scores compared to other approaches at the same time.

A more recent paper employed a U-Net Neural Network architecture [2]. This model simplified upon previous models by regressing the detection model on the probability of pixels being within 4 pixels of the annotated center of a cell. They also employed CNN for cell classification. This indicates to us the for the best results, one should look at implementing a CNN model.

Despite this, our optimized Neural Network Model still gets modest results whilst being far easier to implement and run. Also given the time frame and our machine learning proficiency a Neural Network Implementation is justifiable.

### Ultimate Judgement

Our Ultimate Judgement for the best model for dataset ‘CRCHistoPhenotypes’ would be to implement a CNN model under Deep Learning Framework, in a professional environment where the time and expertise to implement such a model, and hardware which can run such a model are available. Ultimately

The reasoning why CNN would be the best model for dataset ‘CRCHistoPhenotypes’ would be:

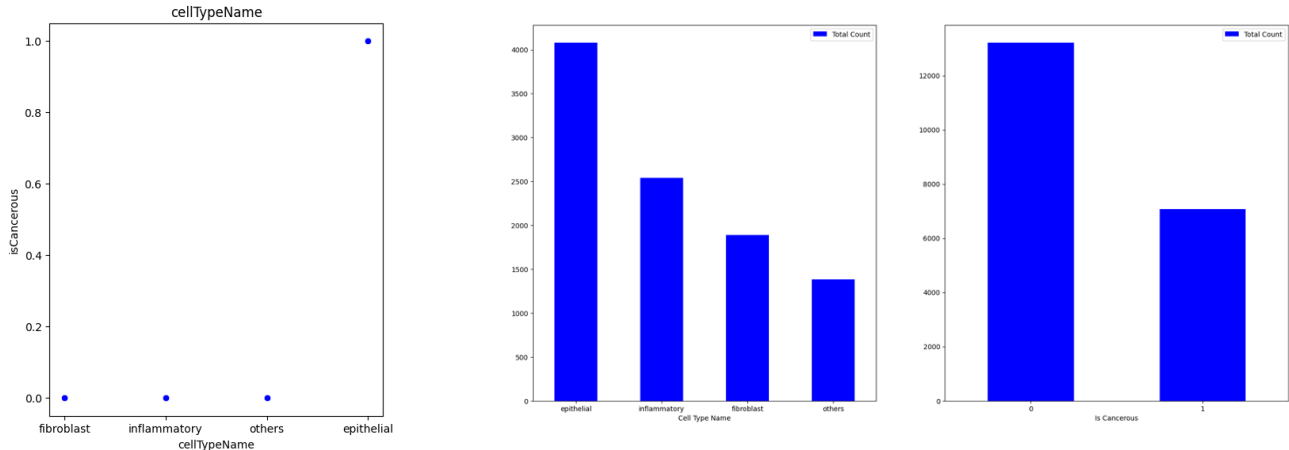
- As previously stated, CNN proved to have the highest F1 scores for the same dataset.
- CNN is more customisable and can be fine-tuned to a further extent than a normal neural network.
- Real world scenario’s are becoming more complex and the solutions to these problems are well suited to CNN.

## Appendices

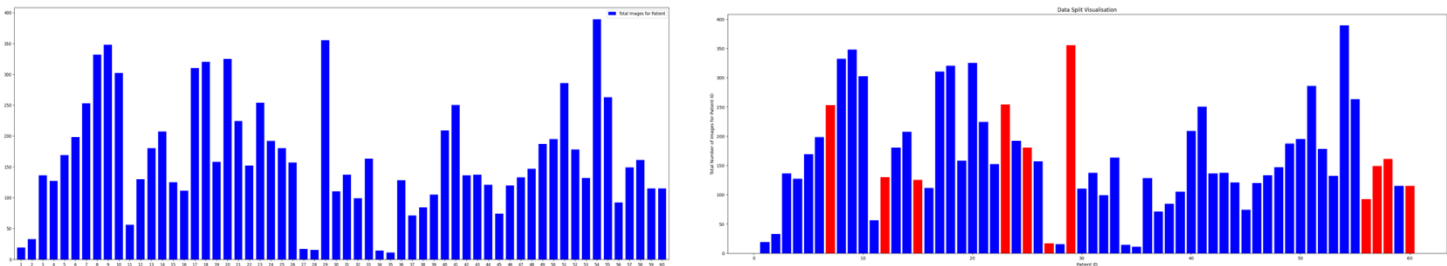
### References:

- [1] <https://ieeexplore-ieee-org.ezproxy.lib.rmit.edu.au/stamp/stamp.jsp?tp=&arnumber=7399414>
- [2] [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3981501](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3981501)

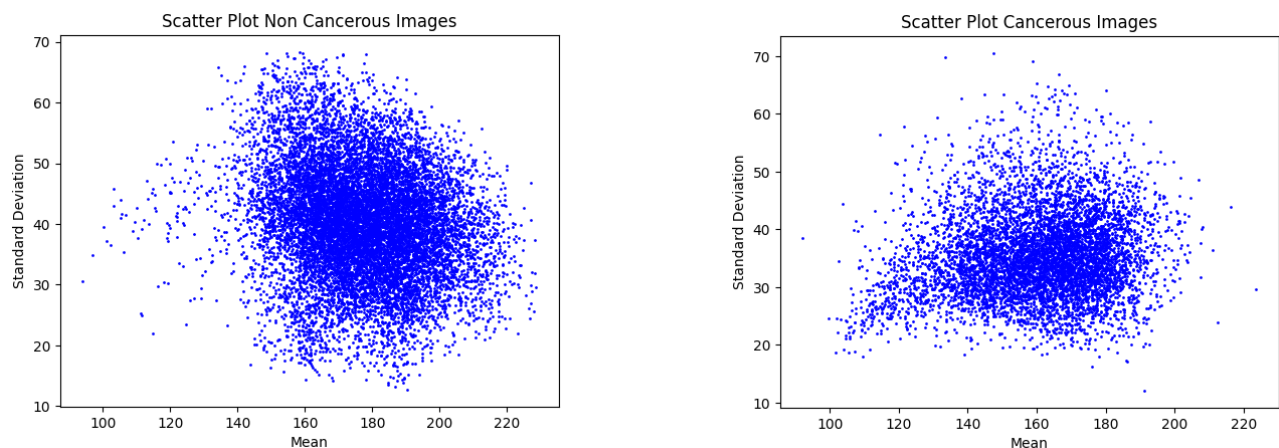
### Appendix 1 – EDA Initial look at raw data



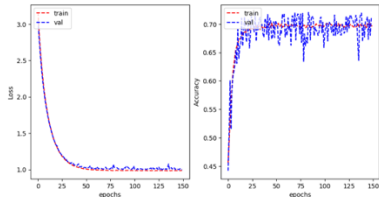
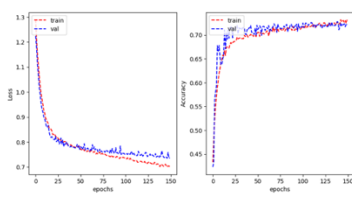
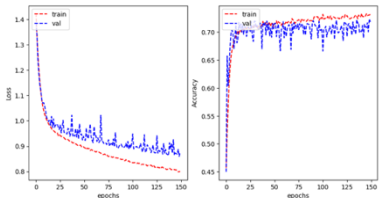
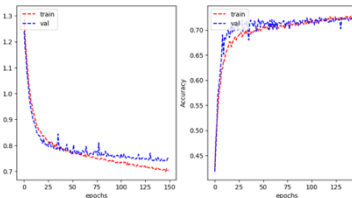
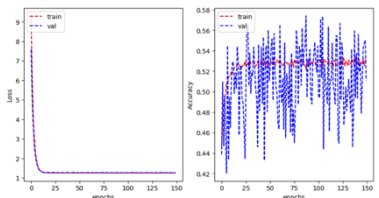
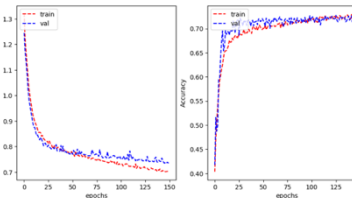
### Appendix 2 – Custom data splitting via Patient ID. Blue = Training/Validation Data Red = Testing Data

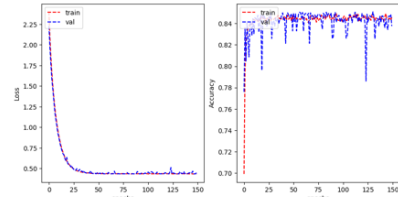
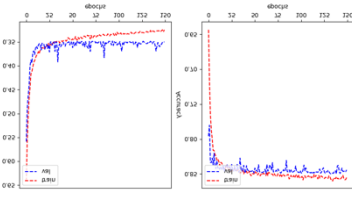
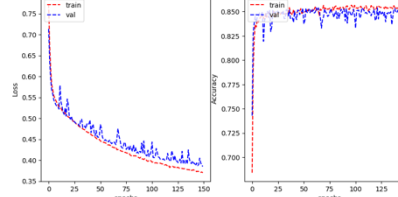
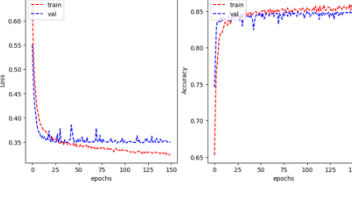
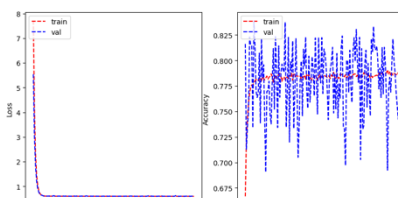
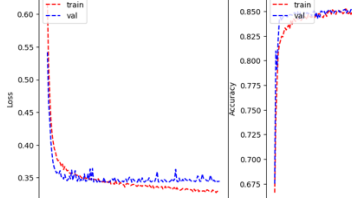


### Appendix 3 – Image scatter plot data



### Appendix 4 – Results of Baseline, Regularization and Dropout optimization, with test data results. For Cell Type Classification and Cancerous Classification

<b>Cell Type Classification</b>	<b>Baseline</b>	Base - Test Loss: 0.9669 Base - Test Accuracy: 0.5970
<b>Regularisation</b>	<b>Drop Out</b>	Reg = Regularisation Drop = Dropout
reg_lambda = 0.01 	reg_lambda = 0.01 	Reg - Test Loss: 1.1345 Reg - Test Accuracy: 0.5924  Drop - Test Loss: 0.9035 Drop - Test Accuracy: 0.6342
reg_lambda = 0.001 	reg_lambda = 0.001 	Reg - Test Loss: 1.0285 Reg - Test Accuracy: 0.6263  Drop - Test Loss: 0.8792 Drop - Test Accuracy: 0.6342
reg_lambda = 0.05 	reg_lambda = 0.05 	Reg - Test Loss: 1.3753 Reg - Test Accuracy: 0.4194  Dropout - Test Loss: 1.6964 Drop - Test Accuracy: 0.2113

<b>Is Cancerous Classification</b>	<b>Baseline</b>	Test Loss: 0.6802 Test Accuracy: 0.6714
<b>Regularisation</b>	<b>Drop Out</b>	Reg = Regularisation Drop = Dropout
reg_lambda = 0.01 	reg_lambda = 0.01 	Reg - Test Loss: 0.4808 Reg - Test Accuracy: 0.8344  Drop - Test Loss: 0.3838 Drop - Test Accuracy: 0.8138
reg_lambda = 0.001 	reg_lambda = 0.001 	Reg - Test Loss: 0.4090 Reg - Test Accuracy: 0.8333  Drop - Test Loss: 0.3754 Drop - Test Accuracy: 0.8268
reg_lambda = 0.05 	reg_lambda = 0.05 	Reg - Test Loss: 0.6245 Reg - Test Accuracy: 0.7028  Drop - Test Loss: 0.3664 Drop - Test Accuracy: 0.8352