COSC2673 – Assignment 2 – Cancerous Cell Image Classification

Name: Nelson Cheng, **Student Number:** S2105802	**Name:** Kah Hie Toh, **Student Number:** S3936897

Task & Approach

The aim of this project is to train two machine learning models for image classification of the provided images of cells to predict whether the cells are cancerous and what type of cell they are. After a process of experimentation, an ultimate judgement will be made of the most recommended model for the tasks. Finally, an independent evaluation will be conducted, comparing the solution with other research and approaches that have been done for similar tasks.

Exploratory data analysis is done initially to gain an understanding of the data and structure, to inform the choices in modelling experimentation. Based on this, a modelling and evaluation framework was selected, including selection of a baseline algorithm, the appropriate evaluation metric and loss function. Next, a baseline model was implemented and evaluated. Further experimentation is then carried out to attempt to improve model performance, analyse results in bias and variance, then reduce any found issues of underfitting or overfitting while managing the model's computational complexity to ensure a reasonable amount of training time. Experimentation includes tuning hyperparameters, different model algorithms, regularization and consideration of extra data. An ultimate judgement will be made on the best models to select for both tasks.

Exploratory Data Analysis

The main dataset contains 9,896 data, all labelled accordingly while the second dataset contains 10,384 data. Although there were no missing values, there were no labels for the cell type in the Extra data, thus some form of semi-supervised learning will be considered. Data shows that cells from a given Patient may be cancerous and of many different cell types, meaning there is no need to split data to prevent data leakage. There are a total of 20,280 colored images combined, indicating that every image is annotated and there are no duplicates.

From **Appendix 1**, [Table 1, Figure 1], the amount of data in the main dataset regarding cancerous cells is fairly balanced. However, the majority of the data contains epithelial cells, with data for other cell types lacking [Table 1, Figure 2]. This could also affect the prediction of cancerous cells as training is done on the same set of images. The number of data for cancerous and non-cancerous cells in the extra dataset is imbalanced too [Table 1, Figure 3]. From these observations, the team could also consider combining the data sets when training on whether the cells are cancerous. Although that does not necessarily help with the dataset being imbalanced [Table 1, Figure 4], the model could train on more data. Additionally, within the main dataset, data with cancerous cells are fairly distributed across 3 different cell types. However, all data of the other class were epithelial cells. [Figure 5] This poses serious issues where the model would have no data of different cell types for non-cancerous cells and vice versa.

Modelling Framework

For image classification tasks, neural networks are said to be one of the most effective solutions due to their ability to generate complex non-linear models. Additionally, Archimbaud (2023) states that Convolutional Neural Networks, CNNs are the most popular and accurate models used for image classification. This is due to the way in which the Convolution layers of a CNN allows the model to extract features from the images to recognise patterns. Therefore, Fully Connected Neural Network and CNN modelling algorithms will be experimented with.

Two optimisation functions namely SGD (Stochastic Gradient Descent) and Adam will be experimented with. SGD is a standard function that updates the weights of connections in the training process, while Adam is a variation of SGD that applies an adaptive learning rate to improve the weighting update process. According to Doshi (2019), Adam is the best optimizer among all others, thus, Adam optimizer will be used on most models.

F1 Score evaluation metric has been selected for both tasks. Accuracy will not be appropriate as the dataset is imbalanced. A badly performed model could always predict the majority class and get a high accuracy. To predict cancerous cells, false positives would be preferred over false negatives when there is a prediction error. The same metric will be used for predicting cell types as it is suitable for imbalanced classification too. (Brownlee 2020)

The activation functions used within the models are ReLU, Sigmoid and Softmax. According to Gupta (2022), Sigmoid is a widely used activation function for binary classification problems. Softmax provides probabilities for each class in

multiclass classification problems. ReLU is used within the layers, as it is a simple and efficient function to allow non-linear modelling. Therefore, when predicting on the isCancerous field, the Sigmoid activation function is used, while Softmax is used for predicting the Cell Type field.

The data is split into 60% Training and 20% for Testing and Validation respectively, ensuring enough data for training and validation, and prediction evaluation is on unseen data from the training process.

Baseline Model

Different models experimented with have been named, with details and results listed in **Appendix 2.** A simple Fully Connected Neural Network model is used for baseline models, with no image preprocessing and the basic SGD optimiser. The Cancerous Binary baseline mode, **B-Baseline**, has predictions with a Training F1 of 0.828 and a Test F1 of 0.891. The F1 performance is good but could be improved upon, with a possible indication of bias in the model. The F1 error gap between the Training accuracy and the Test accuracy is low, meaning this model is generalizing well.

For the Cell Type Multi-class model, the **M-Baseline** predictions had a Training F1 of 0.668 and a Test F1 of 0.764. The F1 performance is low, indicating a strong possibility of bias. Additionally, the model is predicting significantly better on unseen test data. This is an indication that the baseline model is performing poorly.

Model Experimentation

Again, to see model details and results, see **Appendix 2.** The first stage of experimentation attempts to improve accuracy through **increasing the complexity** of the neural network with more layers and neurons. As the complexity increased, the performance on training predictions generally improved. However, the performance of prediction on unseen data did not significantly improve, indicating that bias in the model is decreasing but variance is increasing, likely indicating overfitting. Image preprocessing via **grey-scaling** was also experimented with to reduce training complexity, but it was found that performance was reduced.

Regularization experimentation was then carried out on the 3 Layer NN model. As the 4 Layer model predicted on the training data with a perfect 1.0 rate of accuracy, this indicates that the model is very overfitted, and therefore not the best candidate for improvement. L2 Regularization, Dropout and Early Stopping were tried. The most effective Regularization technique was the **Early Stopping** Method. This method involves stopping the training epochs once the iterations stop yielding improvements on the training/validation error gap (and hopefully overfitting). Early Stopping also has the added benefit of reducing the training time for models, due the the reduction in epoch training iteration. This reduced the training error gap in the IsCancerous model, but did not provide increases in performance, while Regularization was not effective with the Cell Type multiclass modelling.

Data Augmentation is also considered to reduce overfitting. Interestingly, all models trained with augmented data performed worse. As said by Sikka (2020), while the main purpose of data augmentation is to reduce overfitting, some augmentation combinations could cause models to underfit, which seems like the case here. As the provided images visually have little features, performing augmentation might cause some images to look unrecognizable.

Hyperparameter Tuning was then applied to this model to find an optimal initial learning rate and momentum. So far, the best performing NN for the Cancerous Binary task is **B-01,** with a Training F1 of 0.895 and a Test F1 of 0.897. For the Multi-class Cell Type Model, the best performing model is **M-02**, with a Test F1 of 0.884 and a Training F1 of 0.784, which still shows high variance and likely overfitting.

Next, experiments were run with **Convolutional Neural Networks** to attempt to improve on the model's performance. The Convolution filters in CNNs allow them to extract features out of the images to recognise patterns. Similar experimentation was performed on the CNNs through increasing complexity in the convolutions and classifier layers. The convolutional layers were tuned via the channels (i.e. the number of feature maps), kernel sizes, stride and zero padding applied per layer. The classifier layers were tuned with different layers and neurons per layer.

For both tasks, the CNN configurations performed slightly better than the NN configurations. For the Cancerous modelling the best configuration was **B-11** with a Training F1 of 0.933 and a Test F1 of 0.907**.** For the Cell Type modelling the best configuration was **M-09**, with a Training F1 of 0.833 and a Test F1 of 0.797. ROC curves for the Cell Type Predictions can be found in Appendix 1 [Table 2].

The final avenue attempted to improve performance in these models was to add new data in model training. Two methods were investigated to achieve this. The first of which was **Data Augmentation** techniques, which can increase the initial training data by adding variations of training images to the dataset. The second method was to incorporate the Extra data. While the Extra data can easily be added for the isCancerous modelling, incorporating this data for Cell Type modelling required an implementation of **Semi-Supervised Learning**, as it is unlabelled by Cell Type. The process of semi-supervised learning is an iterative method, where a baseline model is trained on the labelled data. This model is then used to predict on the unlabelled data. Predictions made that have a very high confidence score according to the Softmax probability are then "pseudo-labeled" then incorporated back into the training data and a new model is trained. This occurs iteratively until little or no predicted data is considered high confidence.

However, the experimentation of semi-supervised learning performed poorly in the cell-type modelling. This is likely due to the problem of class-imbalance in the total dataset. Data Evaluation shows that in the Main data, 58.8% of the data is Cancerous cells, while in the Extra data 71.2% of the data is Cancerous. Since all cancerous data appears to be the Epithelial cell type, we can infer that the remaining 28.8% of data in the Extra data is shared amongst the other three cell types, meaning the data is skewed to the epithelial cell type. This can cause the Neural Network modelling to be not as effective, as class imbalance can negatively affect Neural Networks (Huang et al., 2022).

Ultimate Judgement

For the Binary IsCancerous Modelling task, despite the **B-11** model had a slightly higher performance (F1 score of 0.907 on Test Data), the recommended model for this task is the **B-13** model, which is a simpler CNN with a **Training F1 of 0.912** and a **Test F1 of 0.899**. The B-13 model has almost as good F1 and has a low amount of variance. The reason for this selection is that the model is much simpler in the convolution and classification layers. While the B-11 model would take multiple hours to train, the B-11 trains much faster, with comparable performance.

With the Multi-class Cell Type Modelling task, the recommended model is the **M-09**, which is the simpler CNN model, with a **Training F1 of 0.833** and a **Test F1 of 0.797**. Additionally, this model is efficient in training time.

Independent Evaluation

The first source reviewed was the provided paper by Sirinukunwattana et. al. 2016. This paper studies the use of a spatially-constrained CNN variant (SC-CNN) for cell type classification. In brief, the team applies grey-scale image preprocessing, then makes use of a CNN architecture that is relatively simple in comparison to the CNN structures used here. However, they also had access to more labelled training data, with 22,444 labeled images, which would be of great benefit. Their best performing model was the "Softmax CNN + NEP" model, achieving a weighted average F1 Score of 0.784. As can be seen, the recommended model from this report performs well in comparison, with an F1 of 0.797.

Alom et al. (2022) performed nuclei classification using their proposed model DCRN, short for Densely connected recurrent convolutional network. Their proposed deep CNN model utilizes the structure of Densely Connected Networks (DCN) where the output from the previous layers are used as input for the subsequent layers, reusing features inside the model to improve its performance. Their models were trained on a total of 16,329 patches, with around 4000 to 5000 patches each for epithelial, fibroblast and inflammatory cells, with the remaining 1390 patches for miscellaneous. The model achieved an average F1-score of 0.811 with an AUC of 96.12%. Although the metrics used for evaluation are different, it is clear that their model outperformed the team's. This could be attributed to the fact that they had a balanced dataset, with enough data for each type of cell whereas the provided dataset did not have a lot of data for cell types such as inflammatory and fibroblast cells.

The final source is "Deep Neural Network Models for Colon Cancer Screening" (Kavitha, M.S. et al. 2022). The report studies impressive models with high levels of performance. The data used was of higher resolution, with one study using images of size 224 x 224 pixels, as opposed to the images here being 27 x 27. Secondly, this higher resolution of images allows for a larger number of features, allowing for more complex CNN architectures like AlexNet, DenseNet and ResNet. This allows for more sophisticated feature extraction and processing. Finally, advanced techniques such as Transfer Learning were used, allowing for better starting models, higher accuracy and faster training. Results quoted from this article have some models with an accuracy of up to 96.98%, showing that these advanced models can perform significantly better than the simpler models in this report.

Appendix 1: Figures

Table 1:

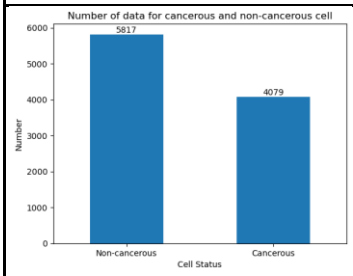
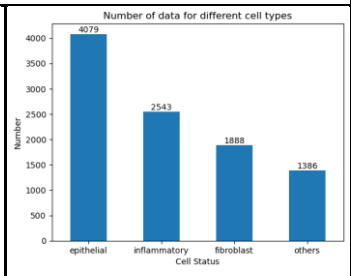
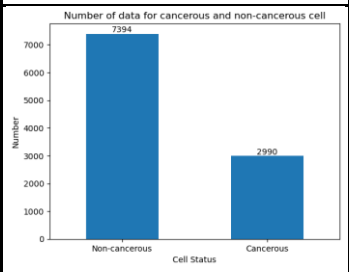
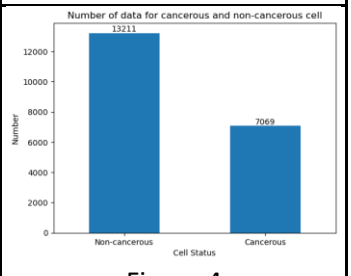
Main Dataset		Extra Dataset	Combined Dataset
 Figure 1	 Figure 2	 Figure 3	 Figure 4

Figure 5:

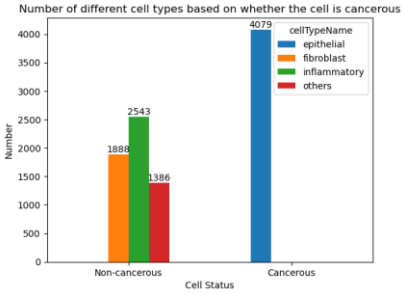
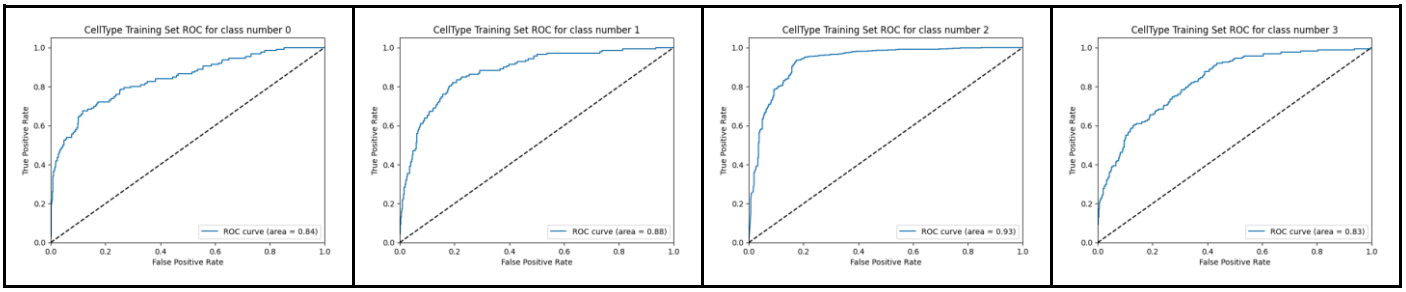


Table 2:



Appendix 2: Model Results

All models use the Adam Optimiser, unless otherwise specified

IsCancerous Model Results

Name	Model Description	Training F1	Test F1
B-Base	Baseline Model - NN, 2 Layers, (729, 729) Neurons, SGD Optimizer	0.828	0.891
B-01	NN, 2 Layers, (729, 729) Neurons	0.895	0.897
B-02	NN, 3 Layers, Neurons (1458, 1458, 729)	0.936	0.883
B-03	NN, 4 Layers, Neurons (1458, 1458, 1094, 729)	1.0	0.854
B-04	NN, 3 Layers, Neurons (1458, 1458, 729), with L2 Regularization	0.977	0.879
B-05	NN, 3 Layers, Neurons (1458, 1458, 729), with Dropout 0.5	0.997	0.878
B-06	NN, 3 Layers, Neurons (1458, 1458, 729), with Early Stopping	0.931	0.891
B-07	NN, augmented data, 2 layers, 256 neurons, Sigmoid Optimizer	0.7388	0.7516
B-08	NN, augmented data, 2 layers, 256 neurons, Sigmoid Optimizer, L2 Regularization	0.6852	0.6965
B-09	NN, augmented data, 2 layers, 256 neurons, Sigmoid Optimizer, L2 Regularization, 0.3 Dropout	0.6970	0.7074
B-10	NN, Grayscale training data, 2 layers, 256 neurons, Relu Optimizer	0.8122	0.7777
B-11	CNN, 3 Conv Layers, Channels (32, 64, 128), Kernels (3, 3, 3), Padding(0, 1, 1). 2x MaxPools, 3 NN Layers, Neurons (4608, 4608, 2306) with Early Stopping	0.933	0.907
B-12	CNN LeNet, 3 Conv Layers, Channels (6, 16, 64), Kernels (5, 5, 3), Padding(0, 1, 1). 3	0.883	0.897

	NN Layers, Neurons (1024, 1024, 512) with Early Stopping		
B-13	CNN, 3 Conv Layers, Channels (64, 128, 256), Kernels (5, 3, 3), Padding(0, 1, 1). 3 NN Layers, Neurons (1024, 1024, 512) with Early Stopping	0.919	0.899

Cell Type Model Results

Name	**Model**	**Training F1**	**Test F1**
M-Base	Baseline Model - 2 Layers, 729 Neurons with SGD optimiser	0.668	0.764
M-01	NN, 2 Layers, (729, 729) Neurons	0.872	0.773
M-02	NN, 3 Layers, Neurons (1458, 1458, 729)	0.884	0.784
M-03	NN, 4 Layers, Neurons (1458, 1458, 1094, 729)	0.97	0.687
M-04	NN, 3 Layers, Neurons (1458, 1458, 729),with L2 Regularization	0.989	0.737
M-05	NN, 3 Layers, Neurons (1458, 1458, 729), with Dropout 0.5	0.844	0.748
M-06	NN, 3 Layers, Neurons (1458, 1458, 729)s with Early Stopping	0.916	0.778
M-07	CNN, 3 Conv Layers, Channels (32, 64, 128), Kernels (3, 3, 3), Padding(0, 1, 1). 2x MaxPools, 3 NN Layers, Neurons (4608, 4608, 2306) with Early Stopping	0.81	0.793
M-08	CNN LeNet, 3 Conv Layers, Channels (6, 16, 64), Kernels (5, 5, 3), Padding(0, 1, 1). 3 NN Layers, Neurons (1024, 1024, 512) with Early Stopping	0.775	0.779
M-09	CNN, 3 Conv Layers, Channels (64, 128, 256), Kernels (5, 3, 3), Padding(0, 1, 1). 3 NN Layers, Neurons (1024, 1024, 512) with Early Stopping	0.833	0.797

Appendix 3: References

- Archimbaud, E. (2023). Programming Image Classification with Machine Learning. [online] Available at: https://kili-technology.com/data-labeling/computer-vision/image-annotation/programming-image-classification-with-machine-learning.
- Gupta, D. (2020). Activation Functions | Fundamentals Of Deep Learning. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2020/01/fundamentals-deep-learning-activation-functions-when-to-use-them/.
- Doshi, S. (2020). Various Optimization Algorithms For Training Neural Network. [online] Medium. Available at: https://towardsdatascience.com/optimizers-for-training-neural-network-59450d71caf6#:~:text=Adam%20is%20the%20best%20optimizers.
- Sikka, M. (2020). Balancing the Regularization Effect of Data Augmentation. [online] Medium. Available at: https://towardsdatascience.com/balancing-the-regularization-effect-of-data-augmentation-eb551be48374.
- Alom, Z., Asari, V.K., Parwani, A. and Taha, T.M. (2022). Microscopic nuclei classification, segmentation, and detection with improved deep convolutional neural networks (DCNN). Diagnostic Pathology, 17(1). doi:https://doi.org/10.1186/s13000-022-01189-5.
- Huang, Z., Sang, Y., Sun, Y. and Jiancheng Lv (2022). A neural network learning algorithm for highly imbalanced data classification. Information Sciences, 612, pp.496–513. doi:https://doi.org/10.1016/j.ins.2022.08.074.
- K. Sirinukunwattana, S. E. A. Raza, Y. Tsang, D. R. J. Snead, I. A. Cree and N. M. Rajpoot (2016), “Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images” in IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1196-1206, May 2016, doi: 10.1109/TMI.2016.2525803.
- Kavitha, M.S. et al. (2022), “Deep neural network models for colon cancer screening, Cancers”. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9367621/ (Accessed: 15 May 2023).
- Brownlee, J. (2020). Tour of Evaluation Metrics for Imbalanced Classification. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/.

Code Repository: https://github.com/s2105802-Nelson/COSC2673-A2-Group42