Question 1 Report

Introduction

Cells and Cancer

Cells are the most critical building blocks of living things, including humans. Various types of cells in the human body make up different types of tissue that form organs. These organs in our body make up different systems with different functions. For instance, the gastrointestinal system helps break down foods into usable or absorbable nutrient molecules to support the normal operation of our body [1]. The proper functioning of these systems relies on the appropriate operation of cells. However, if the cells do not work normally, we may have some issues with our bodies.

Cancer is such a disease of cells and is a worldwide leading cause of death [2]. It refers to the abnormal growth of cancer cells in the tissue, forming benign or malignant tumours. Cancer cells in malignant tumours can invade other parts of the body through the blood or lymphatic system and generate more malignancies [3], threatening life. Through decades of research, scientists developed a range of methods to evaluate and treat cancer, among which genomic analysis, such as the next generation sequencing (NGS), can be used for personalized medicine in cancer treatment [4]. However, such genomic analysis can be influenced by the tumour purity of samples, which is the proportion of cancer cells in the tumour [5]. As explained in this research paper [5], sufficient tumour content is critical for NGS to achieve results with good quality, and the tumour purity gives an idea of the normal cell contents in the tumour, which is also important. However, existing approaches for measuring the tumour purity either involve the manual count of cells by pathologists or is limited by the requirements of samples and the lack of spatial organization information [5]. Thus, developing a new method with limited human involvement to predict the tumour purity cost-effectively using widely available samples efficiently is important and beneficial for understanding cancer and assisting the improvement of treatment. This is also the value of the research done by Oner et al. [5].

Machine Learning Technology

Nowadays, artificial intelligence (AI) is a hot topic and is commonly used in various fields, such as computer science and biology. It refers to the ability of the machine to perform tasks by replicating the human cognitive, intellectual process [6]. Machine learning (ML) is a sub-branch of AI that aims to train the machine to learn to perform tasks. However, some basic classifiers used as supervised ML technologies with labelled data requires the time-consuming feature extraction process to select the most relevant and important features for training [7]. In terms of unsupervised ML, like clustering, there are also drawbacks. For example, clustering is computationally expensive and may not be efficient for big data analysis. More importantly, these technologies may not be suitable for predicting tumour purity.

More recently, a sub-branch of ML, deep learning (DL), has become more and more popular. By mimicking the biological neural networks, DL neural networks are comprised of neurons (nodes) and connections between them with weights and bias. Deep neural networks have more than one hidden layer between the input and output layer and can learn the complex or latent patterns in the data without a feature selection process. Additionally, the architecture of the deep learning neural network can be varied to fit different tasks. With these strengths, DL technologies are widely used in biomedical areas, such as imaging. Based on my experience of

using DL neural networks to model the functional architecture of the human brain using functional magnetic resonance imaging (fMRI) data, I found DL is a potent tool to learn features automatically and may unveil some latent things from the data. Thus, I think using neural networks to predict tumour purity is the most suitable. In this paper, Oner et al. [5] used ResNet18 and multi-layer perceptron (MLP) to extract features and predict.

Methods

Model Architecture

After reading the research paper [5], I implemented a similar ML model and used the MNIST digits dataset for training and testing. The components of the model is the same as the one used by Oner et al. To make sure I implement the model with the same architecture as the one used in the paper [5], I adapted the source code published by Oner to suit the use of the different dataset. The architecture of the model is shown in *Figure 1*. The code with more specific architecture design can be found in GitHub.

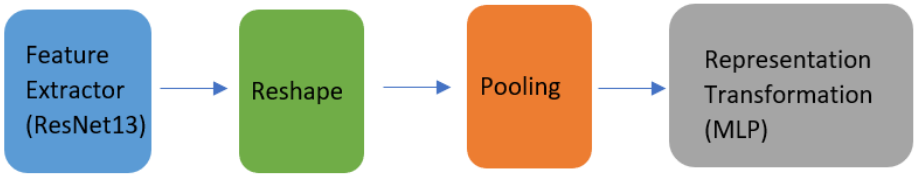

Figure 1 Model Architecture

Data

The data I used for training and evaluating the model is from the MNIST digit dataset [8]. Only the images of zero or seven were used. These data are loaded with a patch size of 100. The tailing images were discarded. The training data were shuffled, but the data for validation were not shuffled.

Training and Validation

After loading the data and get the model prepared, I trained the data using 30 epochs. The loss functions and optimizer used are also the same as Oner et al. [5] used, which are the L1loss from torch that measures the mean absolute error and the Adam optimizer with a 1e-4 learning rate and a 0.0005 weight decay, respectively. The loss calculated during training and validation were recorded and plotted. More parameters can be found in the code loaded on my GitHub repository

Results and Discussion

After training with 30 epochs, I got the plot of the loss during training and validation. As shown in *Figure 2*, although there were some fluctuations in loss after epoch 6, the model seems to converge after 10 epochs and have a loss around 3.47. This shows that the model successfully learned to classify between zero and seven hand written digits.

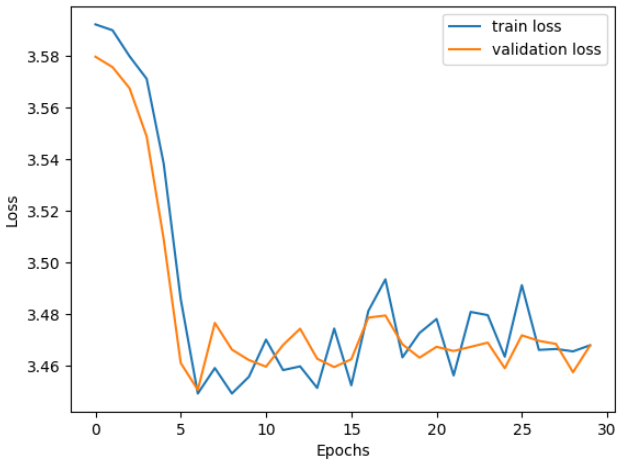

Figure 2 Loss During Training and Validation

Conclusion

As I am doing research in brain functions, I have limited knowledge of cancer. However, through reading the research paper [5] about predicting tumour purity and doing some literature reviews on cancer, I learned a lot about cancer and tumour purity. I understood that cancer is worth researching, and tumour purity is critical for personalized treatment. This research area is fascinating, and I would like to learn more about it. Additionally, implementing the DL model allowed me to practice the coding of generating, training, and evaluating ML models. Through this practice, I am more convinced that I am interested in biomedical data science and am passionate about learning more to develop and improve my ability to use artificial intelligence to solve biomedical problems in practice.

References

[1] N. G. MacFarlane, "Digestion and absorption," *Anaesthesia & Intensive Care Medicine,* vol. 19, no. 3, pp. 125-127, 2018.

[2] M. Uhlen *et al.*, "A pathology atlas of the human cancer transcriptome," *Science,* vol. 357, no. 6352, 2017.

[3] L. A. Liotta and E. C. Kohn, "Cancer's deadly signature," *Nature genetics,* vol. 33, no. 1, pp. 10-11, 2003.

[4] S. Morganti, P. Tarantino, E. Ferraro, P. D'Amico, B. A. Duso, and G. Curigliano, "Next Generation Sequencing (NGS): a revolutionary technology in pharmacogenomics and personalized medicine in cancer," *Translational Research and Onco-Omics Applications in the Era of Cancer Personal Genomics,* pp. 9-30, 2019.

[5] M. U. Oner *et al.*, "Obtaining spatially resolved tumor purity maps using deep multiple instance learning in a pan-cancer study," *Patterns,* p. 100399, 2021.

[6] S. L. Goldenberg, G. Nir, and S. E. Salcudean, "A new era: artificial intelligence and machine learning in prostate cancer," *Nature Reviews Urology,* vol. 16, no. 7, pp. 391-403, 2019.

[7] S. A. Bini, "Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care?," *The Journal of arthroplasty,* vol. 33, no. 8, pp. 2358-2361, 2018.

[8] Y. LeCun *et al.*, "Comparison of learning algorithms for handwritten digit recognition," in *International conference on artificial neural networks*, 1995, vol. 60: Perth, Australia, pp. 53-60.