Deep Learning

880663-M-6

Assignment

Using Deep Learning to Perform Multi-Class Classification on the Lung and Colon Cancer Histopathological Image Dataset (LC25000)

Report by:

Robert Dekkers (Anr. 545046)

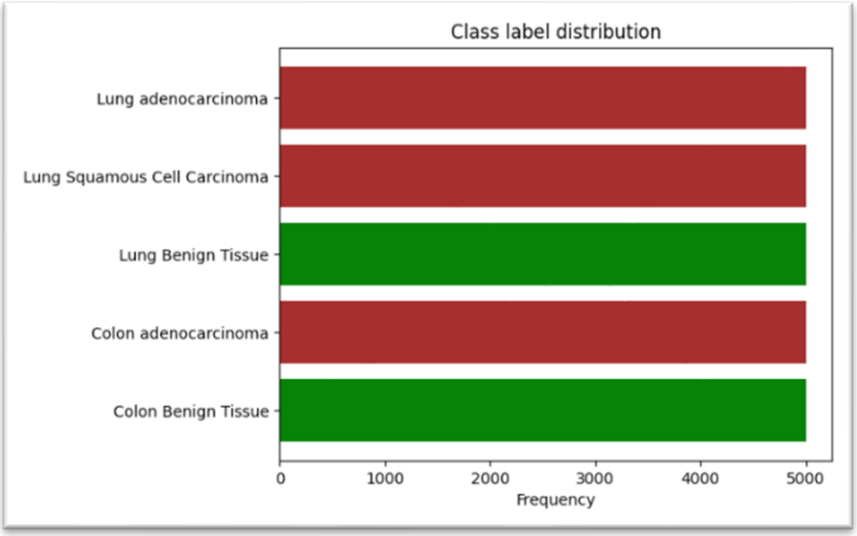
March 2024

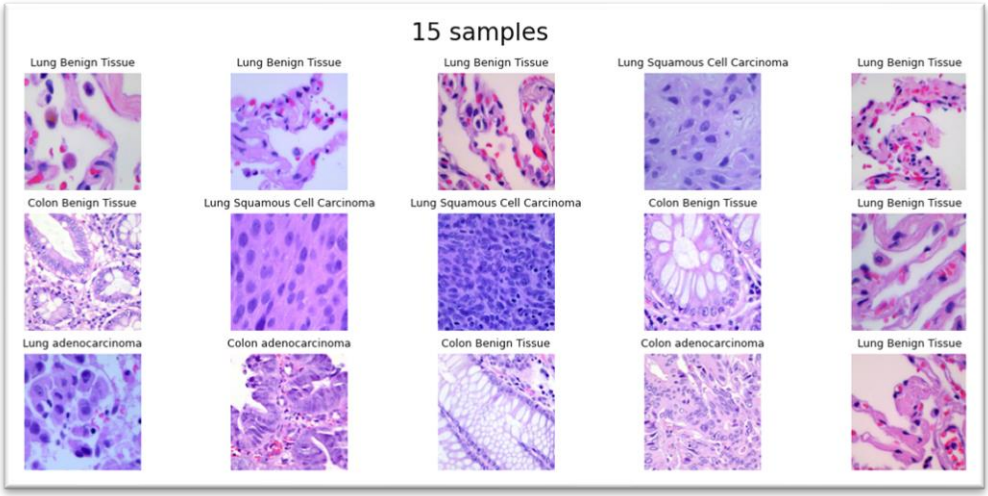
1. Problem Definition

Histopathology is the study of disease at a biological tissue level. The Lung and Colon Cancer Histopathological Image Dataset (LC25000) is created to facilitate the development of machine learning methods for cancer diagnosis (Borkowski et al., 2019). In this assignment, the objective is to correctly classify into one of the five categories of images.

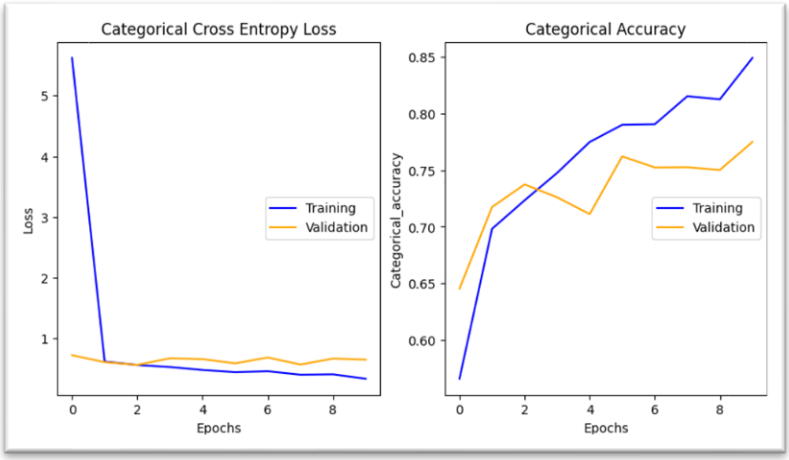
2. Exploratory Data Analysis

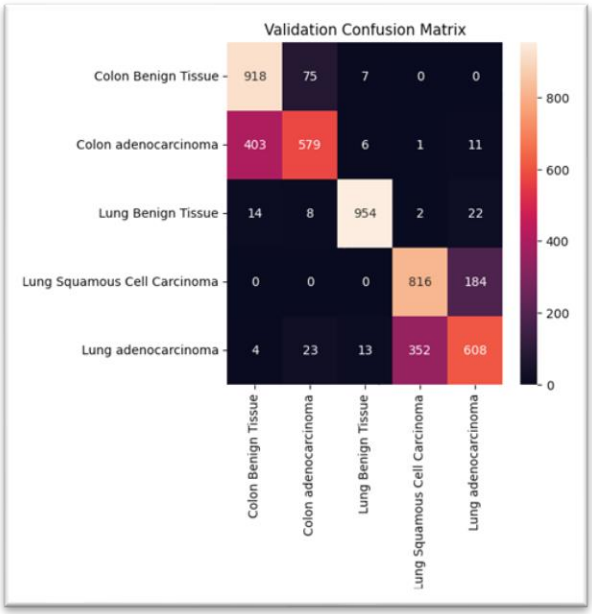
The dataset contains 25000 images equally distributed into 5 classes of 5,000 images, containing benign- and cancerous tissues, and lung- and colon tissues. For binary cancer prediction, this dataset would suffer from class imbalance (10,000 benign images to 15,000 cancer images), but for 5-class classification the dataset is perfectly balanced.

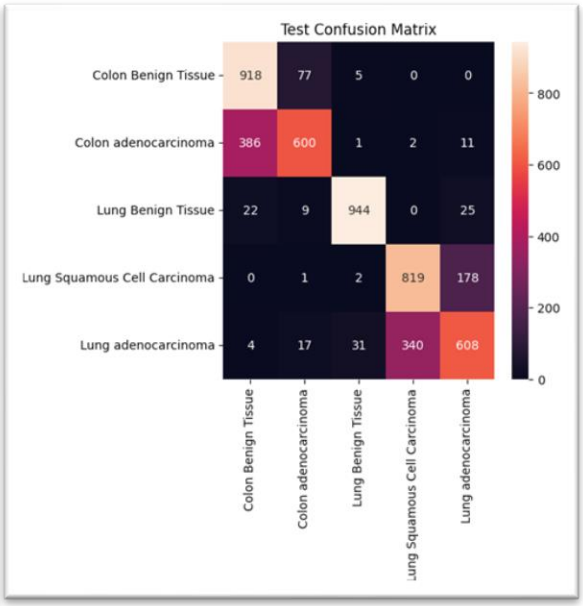




3. Results of the Baseline Model





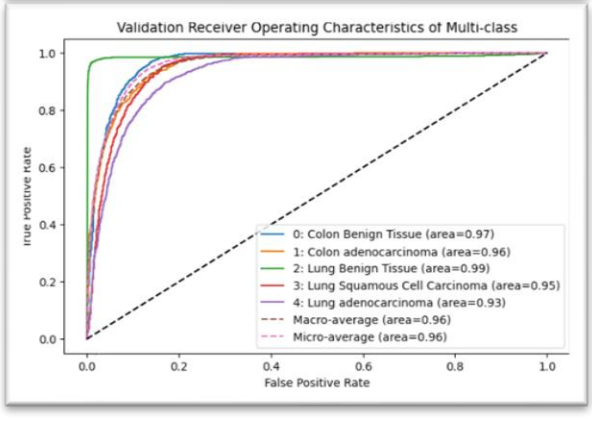


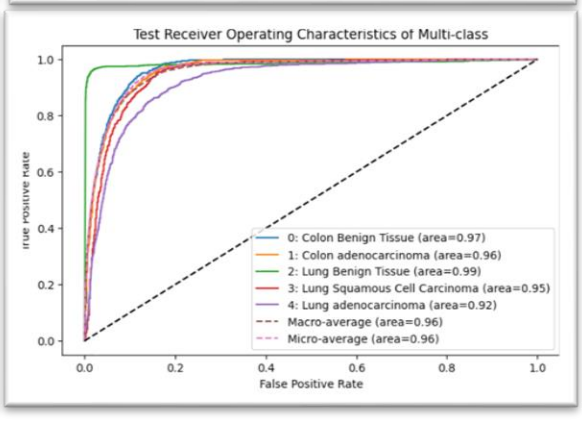
Classification Report

	precision	recall	f1-score	support
Colon Benign Tissue	0.69	0.92	0.78	1000
Colon adenocarcinoma	0.85	0.58	0.69	1000
Lung Benign Tissue	0.97	0.95	0.96	1000
Lung Squamous Cell Carcinoma	0.70	0.82	0.75	1000
Lung adenocarcinoma	0.74	0.61	0.67	1000
accuracy			0.78	5000
macro avg	0.79	0.77	0.77	5000
weighted avg	0.79	0.78	0.77	5000

Classification Report

	precision	recall	f1-score	support
Colon Benign Tissue	0.69	0.92	0.79	1000
Colon adenocarcinoma	0.85	0.60	0.70	1000
Lung Benign Tissue	0.96	0.94	0.95	1000
Lung Squamous Cell Carcinoma	0.71	0.82	0.76	1000
Lung adenocarcinoma	0.74	0.61	0.67	1000
accuracy			0.78	5000
macro avg	0.79	0.78	0.77	5000
weighted avg	0.79	0.78	0.77	5000





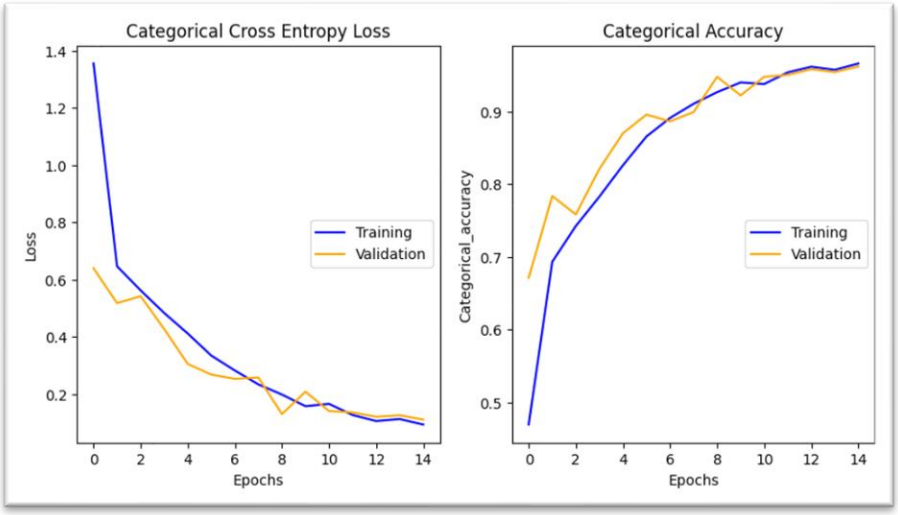
4. Improved (Fine-tuned) Model and Its Results

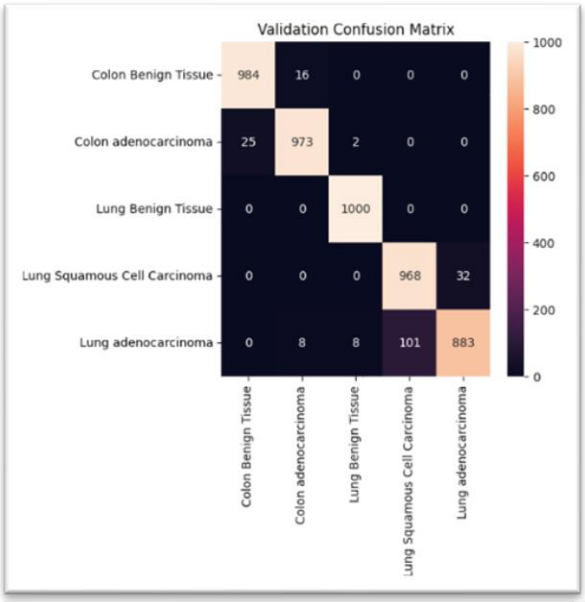
After training the model for 10 epochs, there seems to be no point of diminishing returns yet in terms of validation performance. In other words, there is no sign of overfitting yet, which suggests that the model performance may yet be improved by fitting the model more closely to the data. This may be achieved by training the model for more epochs, or by increasing model complexity, such as by increasing the amount of layers or the amount of units per layer. Below, every experiment is listed in sequence.

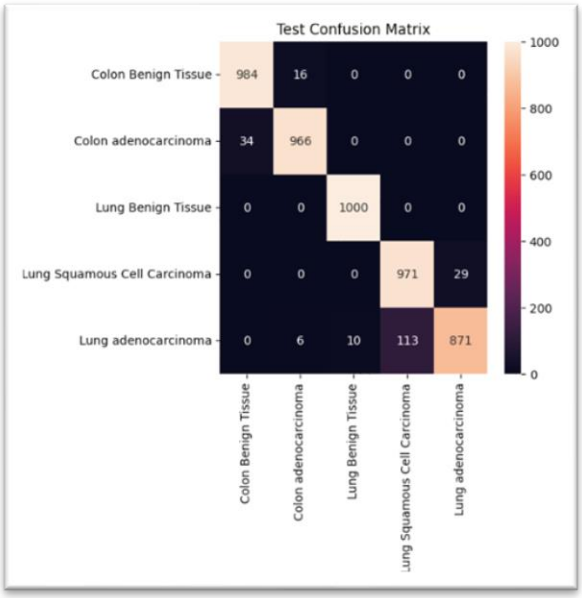
1. The **number of filters** in the second convolution layers is increased from 64 to 256, to add more model complexity. This idea takes inspiration from the VGG16 architecture in which successive layers contain more filters (Simonyan & Zisserman, 2015). Train- and validation accuracy improved.
2. A **convolution layer and max pooling layer are added** after the existing ones with 512 filters, to test whether a more complex model will improve performance. Train- and validation accuracy decreased after 7 epochs.
3. **2 by 2 stride is added** to all convolution layers as means of regularization. Train- and validation accuracy improve beyond previous improvements. Misclassification of images among the same tissue type (lung or colon) notably decreases.
4. The **max pooling layers are replaced by average pooling layers** to see if more information is maintained for a better model fit. The train- and validation loss now decreases more gradually per epoch, instead of immediately during the first epochs. The train- and validation accuracy improve up until the last epoch, without sign of decrease.
5. The **amount of epochs is increased** to 15 to see if the model can fit even better. In the additional 5 epochs, the training accuracy continues to improve, but the validation accuracy decreases slightly, hinting at potential overfitting.
6. **Dropout layers (rate = 0.2) are added** in the fully-connected part of the network with the purpose of improving the generalization performance of the model (Srivastava et al., 2014). The validation accuracy no longer shows signs of diminishing return due to overfitting.

On the next page are the evaluation metrics and diagrams for the final model.

Evaluation results of enhanced model after all experiments





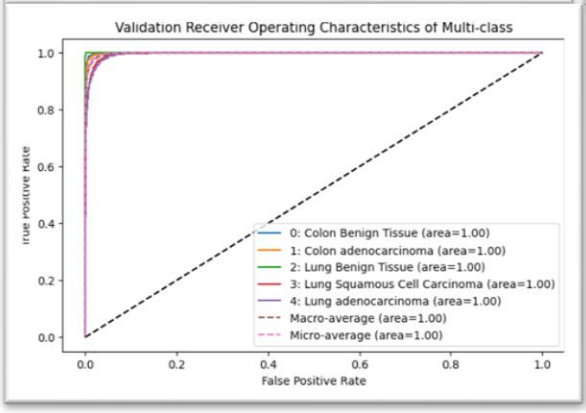


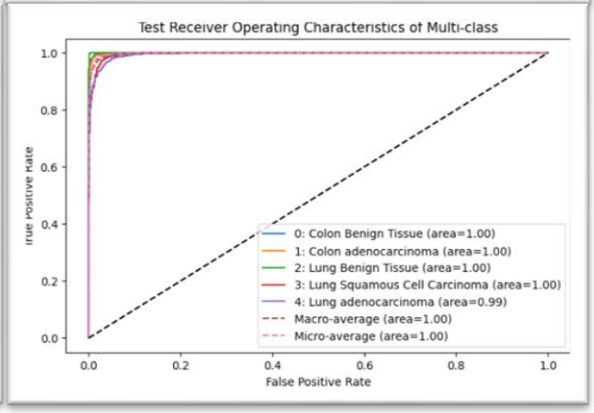
Classification Report

	precision	recall	f1-score	support
Colon Benign Tissue	0.98	0.98	0.98	1000
Colon adenocarcinoma	0.98	0.97	0.97	1000
Lung Benign Tissue	0.99	1.00	1.00	1000
Lung Squamous Cell Carcinoma	0.91	0.97	0.94	1000
Lung adenocarcinoma	0.97	0.88	0.92	1000
accuracy			0.96	5000
macro avg	0.96	0.96	0.96	5000
weighted avg	0.96	0.96	0.96	5000

Classification Report

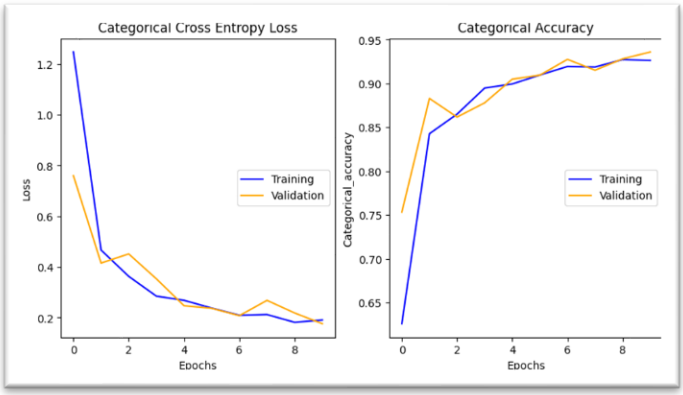
	precision	recall	f1-score	support
Colon Benign Tissue	0.97	0.98	0.98	1000
Colon adenocarcinoma	0.98	0.97	0.97	1000
Lung Benign Tissue	0.99	1.00	1.00	1000
Lung Squamous Cell Carcinoma	0.90	0.97	0.93	1000
Lung adenocarcinoma	0.97	0.87	0.92	1000
accuracy			0.96	5000
macro avg	0.96	0.96	0.96	5000
weighted avg	0.96	0.96	0.96	5000

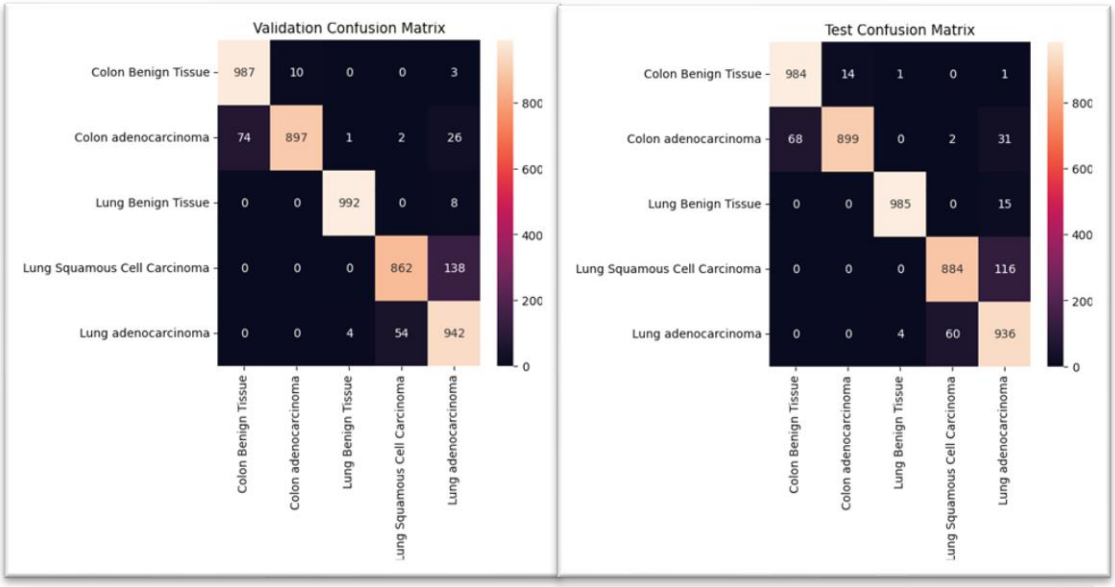




5. Transfer Learning Model and Its Results

A DenseNet121 model was selected for its relatively small size. The complete model architecture has a size in a similar order of magnitude as the baseline model, with 1,184,069 parameters (compared to 745,597). The model performs similar to the enhanced model, with a final test accuracy of 0.94 (compared to 0.96 for the enhanced model).



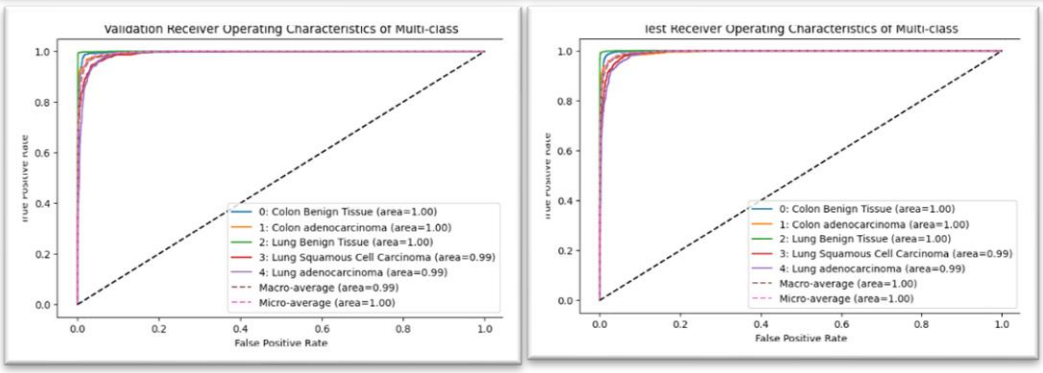


Classification Report

	precision	recall	f1-score	support
Colon Benign Tissue	0.93	0.99	0.96	1000
Colon adenocarcinoma	0.99	0.90	0.94	1000
Lung Benign Tissue	0.99	0.99	0.99	1000
Lung Squamous Cell Carcinoma	0.94	0.86	0.90	1000
Lung adenocarcinoma	0.84	0.94	0.89	1000
accuracy			0.94	5000
macro avg	0.94	0.94	0.94	5000
weighted avg	0.94	0.94	0.94	5000

Classification Report

	precision	recall	f1-score	support
Colon Benign Tissue	0.93	0.99	0.96	1000
Colon adenocarcinoma	0.99	0.90	0.94	1000
Lung Benign Tissue	0.99	0.99	0.99	1000
Lung Squamous Cell Carcinoma	0.94	0.86	0.90	1000
Lung adenocarcinoma	0.84	0.94	0.89	1000
accuracy			0.94	5000
macro avg	0.94	0.94	0.94	5000
weighted avg	0.94	0.94	0.94	5000



6. Discussion

Compared to the baseline model, with a weighted average test accuracy of 0.77, the enhanced model notably improved with a weighted average test accuracy of 0.96. The baseline model performed differently per class, and was exceptionally good at classifying benign lung tissue (f1-score = 0.95) and poor at classifying lung adenocarcinoma tissue (f1-score = 0.67). The enhanced model performed with f1-scores of 1.00 and 0.92 for these classes respectively, showing less difference in performance between classes.

The enhanced model does not show a notable decrease in validation accuracy yet after 15 epochs of training, suggesting that the performance may still be improved further without overfitting. Some experiments that have been run could be further explored by trying out more possible values for the hyperparameters. It is plausible that adding more layers may give a beneficial result, given that the architecture is still relatively shallow compared to architectures such as of VGG16 (Simonyan & Zisserman, 2015) or DenseNet121 (Huang et al., 2018).

Switching and/or tuning the optimizer may also be worthy of exploring. In one experiment, the 'adam' optimizer was replaced by the 'rmsprop' optimizer, but this yielded a poor result by itself. The experiment was replaced by another, and not reported here. The authors of the following cited paper concluded that tuning an optimizer may yield as much result as choosing a different one (Schmidt et al., 2021).

7. References

Borkowski, A. A., Bui, M. M., Thomas, L. B., Wilson, C. P., DeLand, L. A., & Mastorides, S. M. (2019). *Lung and Colon Cancer Histopathological Image Dataset (LC25000)* (arXiv:1912.12142; Version 1). arXiv. https://doi.org/10.48550/arXiv.1912.12142

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2018). *Densely Connected Convolutional Networks* (arXiv:1608.06993). arXiv. https://doi.org/10.48550/arXiv.1608.06993

Schmidt, R. M., Schneider, F., & Hennig, P. (2021). *Descending through a Crowded Valley—Benchmarking Deep Learning Optimizers* (arXiv:2007.01547). arXiv. https://doi.org/10.48550/arXiv.2007.01547

Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition* (arXiv:1409.1556). arXiv. https://doi.org/10.48550/arXiv.1409.1556

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, *15*(56), 1929–1958.