2024

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:

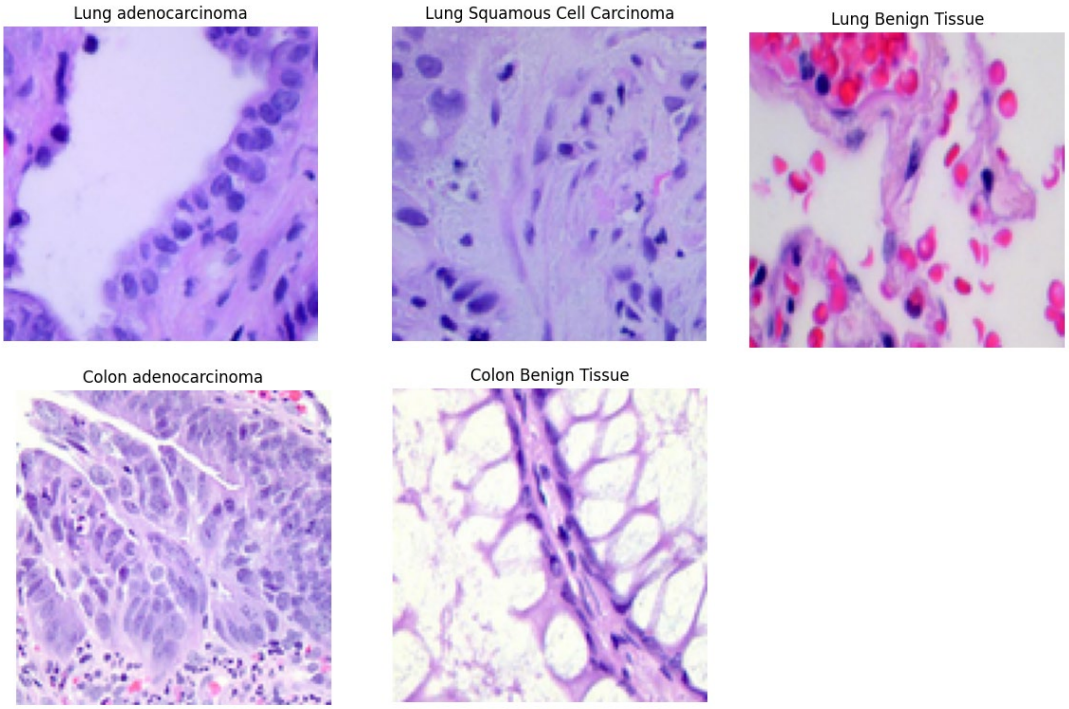
Meliz Tyurkileri (SNR: 2113955)

March 2024

1. Problem Definition

This project is conducted to develop a convolutional neural network will be used in a multi-class classification problem aiming to classifying histopathological images with the Lung and Colon tissues into the following 5 classes (see Figure 1): (i) Lung adenocarcinoma, (ii) Lung squamous cell carcinoma, (iii) Lung benign tissue, (iv) Colon adenocarcinoma, (v) Colon benign tissue.

Figure 1. Histopathological images of lung and colon tissues from LC2500 dataset

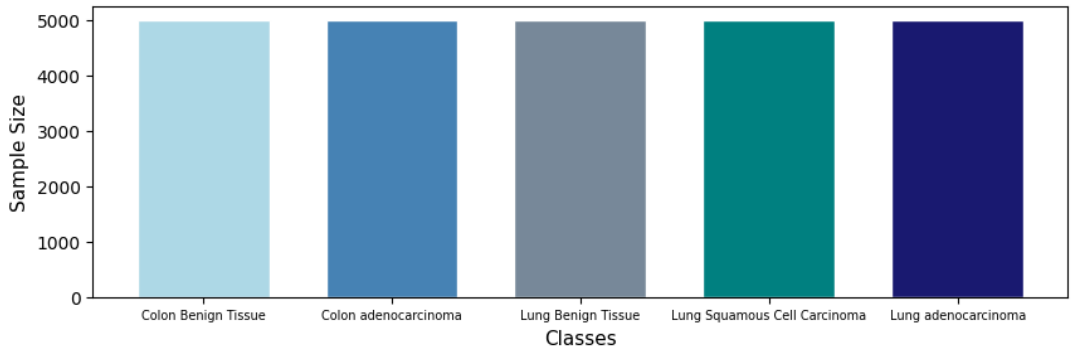


LC2500 dataset used in this study, were generated from the original sample of HIPAA complaints, and the original 1250 images were augmented to 25000 to create LC2500 dataset.

2. Exploratory Data Analysis

As the first step, images were resized from 768 x 768 pixels to 120 x 120 pixels and one hot coding was applied to binarize target values. Final sample includes 25000 images, corresponding to 5000 observations in each classes (see Figure 2). Normalization was not applied in this step considering the memory constraints.

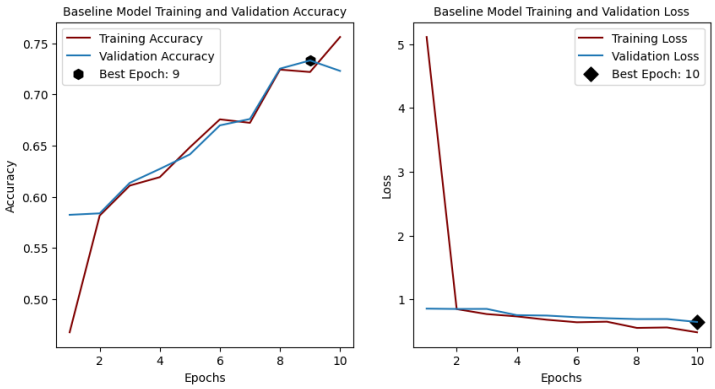
Figure 2. Class distribution



3. Results of the Baseline Model

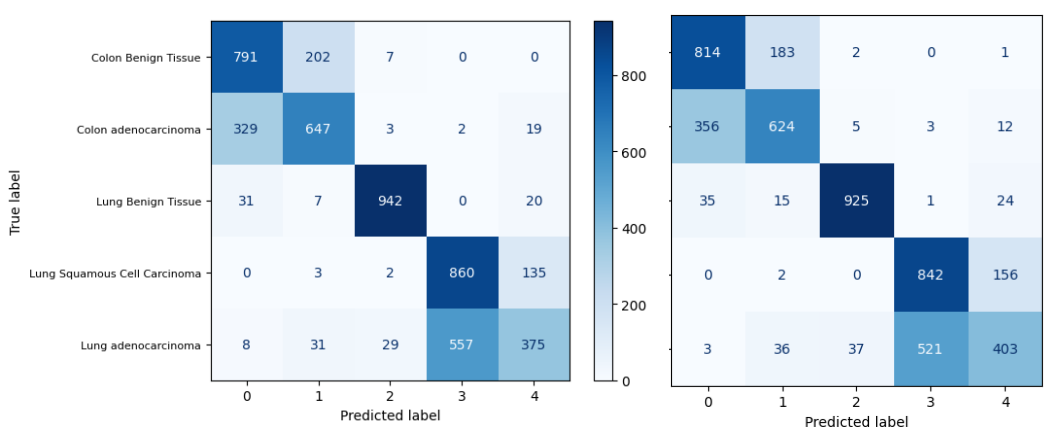
Baseline CNN model was constructed by following the assignment instructions, and the categorical cross entropy was chosen as the loss function. The validation accuracy closely followed training accuracy until the final epoch and started to decline after reaching its peak value, 73% recorded in the 9th epoch (signaling an overfit). On the other hand the gap between training and validation loss was slightly increased following the 4th epoch, while the validation accuracy keep going to improve until the final epoch. Hyperparameters from the best epoch (10) with the lowest validation loss was exploited for the model evaluation and predictions.

Figure 3. Baseline model loss and accuracy across epochs



Both the validation and test accuracies of the baseline model were recorded as 72%. The confusion matrixes and performance measures indicate a clear asymmetries in the succession of the model for segmenting different classes. Indeed, 95% of the figures classified as lung benign tissue in the test data were actually belonged to this group, whereas corresponding value was only 0.62% for lung squamous cell carcinoma (see Figure 4).

Figure 4. Confusion matrix and performance measures for validation and test (baseline model)



Validation

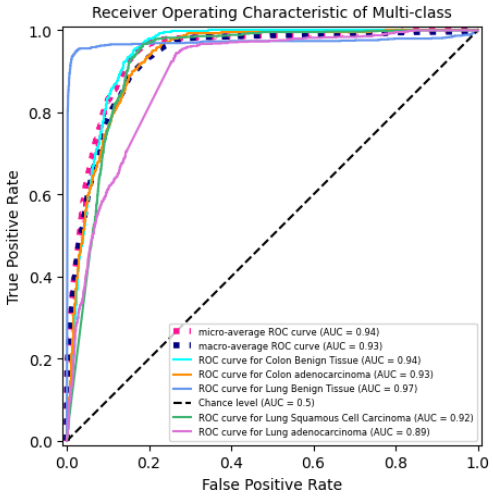
	precision	recall	F1-score	support
Colon benign tissue	0.68	0.79	0.73	1000
Colon adenocarcinoma	0.73	0.65	0.68	1000
Lung benign tissue	0.96	0.94	0.95	1000
Lung squamous cell carcinoma	0.61	0.86	0.71	1000
Lung adenocarcinoma	0.68	0.38	0.48	1000
accuracy			0.72	5000
macro avg	0.73	0.72	0.71	5000
weighted avg	0.73	0.72	0.71	5000

Test

	precision	recall	F1-score	support
Colon benign tissue	0.67	0.81	0.74	1000
Colon adenocarcinoma	0.73	0.62	0.67	1000
Lung benign tissue	0.95	0.93	0.94	1000
Lung squamous cell carcinoma	0.62	0.84	0.71	1000
Lung adenocarcinoma	0.68	0.40	0.51	1000
accuracy			0.72	5000
macro avg	0.73	0.72	0.71	5000
weighted avg	0.73	0.72	0.71	5000

As seen in Figure 5, performance of the baseline model varies across the different classes. While the Lung benign tissue had the best outcome, the performance of the model significantly lagged behind for the lung adenocarcinoma compared to its counterparts.

Figure 5. RoC curve for the test data (baseline)



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

The baseline model was improved using RandomSearch tuner[1]. Before starting the experiment and hyperparameter optimization, one more convolutional layer has been added to the baseline model similar to Masud et al. (2021) and Mangal, Chaurasia and Khajanchi (2020)[2]. 10 different trials was conducted to select 11 distinct hyperparameters as described below. Epochs for each trial were limited to 10 for simplicity, while the batch size was set to 32. Optimization has been conducted within two steps. At the first step, model design and main hyperparameters has been selected via trials, and then the enhanced model was retrained on the training set using larger number of epochs and batch size to finetune weights.

Table 2: Properties of the trials and corresponding validation accuracies

	#8	#7	#00	#4	#6	#9	#3	#2	#5	#1
Conv2D layer 1 filter	128	128	128	128	64	128	64	64	64	128
Conv2D layer 1 kernel	3	5	3	3	3	5	3	3	3	5
Conv2D layer 2 filter	256	128	256	256	512	512	512	128	256	128
Conv2D layer 2 kernel	3	5	3	3	5	3	5	3	3	5
Dropout rate layer 2	0.2	0.1	0.3	0.1	0.2	0.3	0.1	0.3	0.2	0.1
Conv2D layer 3 filter	512	512	256	128	128	128	128	512	128	256
Conv2D layer 3 kernel	5	5	5	3	3	5	5	5	5	3
Dropout rate layer 3	0.1	0.2	0.1	0.3	0.3	0.2	0.1	0.1	0.3	0.1
Dense layer 1 unit	64	64	128	128	256	64	128	64	256	128
Dense layer 2 unit	32	128	32	128	32	128	128	32	128	128
Learning rate	1e-4	1e-4	1e-4	1e-3	1e-4	1e-4	1e-3	1e-4	1e-3	1e-3
Validation Accuracy	0.972	0.969	0.968	0.967	0.965	0.963	0.960	0.958	0.944	0.943

Best model (#8) from the trials reached to 97% validation accuracy (see Table 2), corresponding hyperparameters were derived from the model and it was retrained on the training set using 32 epochs and the batch size of 64. Despite the computational costs and possible overfit, number of epochs was increased in the enhanced model taking into account that the larger epoch size may contribute to the model generalization. On the other hand, early stopping was introduced to prevent a possible

[1] For further details on KerasTuner see https://keras.io/keras_tuner/.

[2] I have also experimented the models with 2 convolutional layers, however, the average validation score remained between 90% and 95%.

overfit, dropout layers has been also added for a similar purpose. As seen in Table 2, 0.001 (default) and 0.0001 (Mangal, Chaurasia & Khajanchi, 2020) learning rates have been randomly tested, and the later outperformed. Although the smaller learning rate slowed down the convergence, it improved the stability and mitigated the risk of overshooting. To speed up the process, I have also introduced the batchMinimization (see Masud et al., 2021; Masud et al., 2020), however, it resulted in higher volatility across the epochs, so I decided to keep the model relatively simple. Adam optimizer was selected as the compiler optimizer, although the current literature refer to the success of another optimizer: RMSprop (see Masud et al., 2021; Masud et al., 2020; Mangal, Chaurasia & Khajanchi, 2020) its performance is dependent on the given parameters and demands more precise determination of the optimizer parameters.

Table 2: Properties of the selected enhanced model

1st Convolution layer	Filter size = 128, kernel_size =3, activation= Relu, zero padding
Max pooling	3x3
2nd convolution layer	Filter size = 256, kernel_size = 3, activation= Relu, zero padding
Max pooling	3x3
Drop out – 2nd layer	Rate = 0.2
3rd convolution layer	Filter size = 512, kernel_size = 5, activation= Relu, zero padding
Max pooling	3x3
Drop out – 3rd layer	Rate = 0.1
Flatten	
1st Dense layer	Units = 64
2nd Dense layer	Units = 32
Dense layer activation	Softmax, 5 units output layer
Compiler optimizer	Adam with 1e-4 learning rate
Compiler loss	Categorical crossentropy
Batch size & epoch	64 batch size, 32 epochs with early stopping (patience=5, monitor=validation loss)

Figure 6. Enhanced model loss and accuracy across epochs

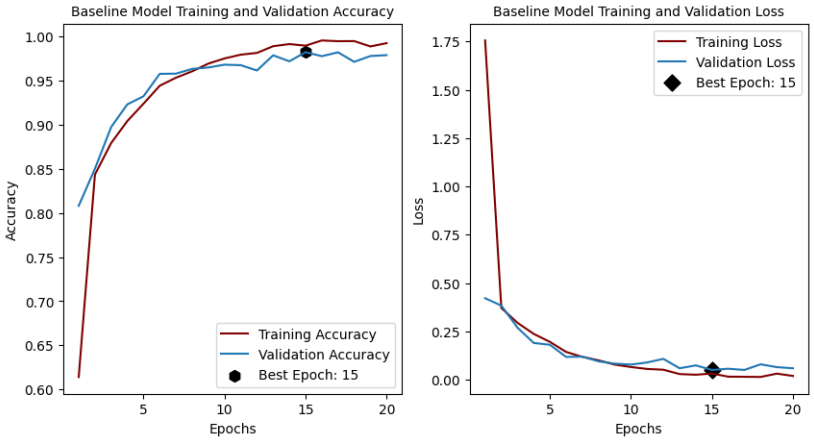
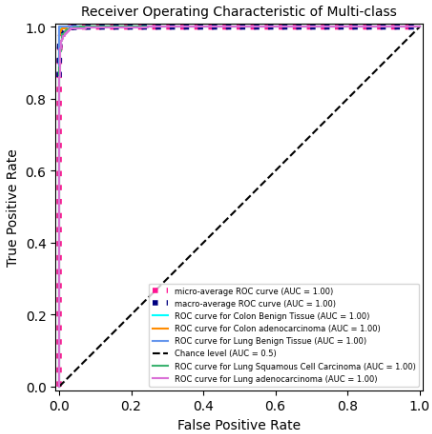
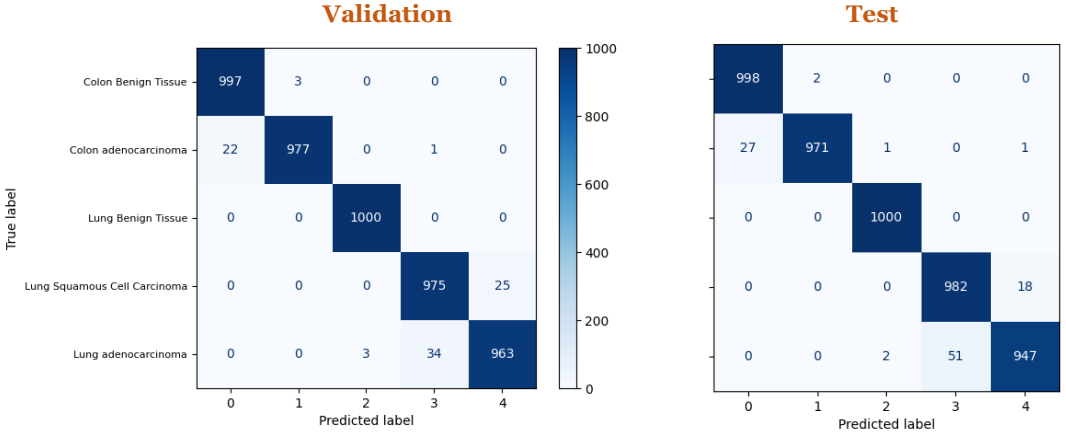


Figure 7. RoC curve for the test data (enhanced model)



As seen in Figure 6, validation accuracy exceeded 95% after the 5th epoch and keep going to improve until the 15th epoch, the minimum loss has been also recorded in this epoch, while it stopped to improve afterwards. Overall, the enhanced model produced a more balanced performance to segment different classes (Figure 7) and reached to 98% average accuracy on the test data (Figure 8).

Figure 8. Confusion matrix and performance measures for validation and test (enhanced model)

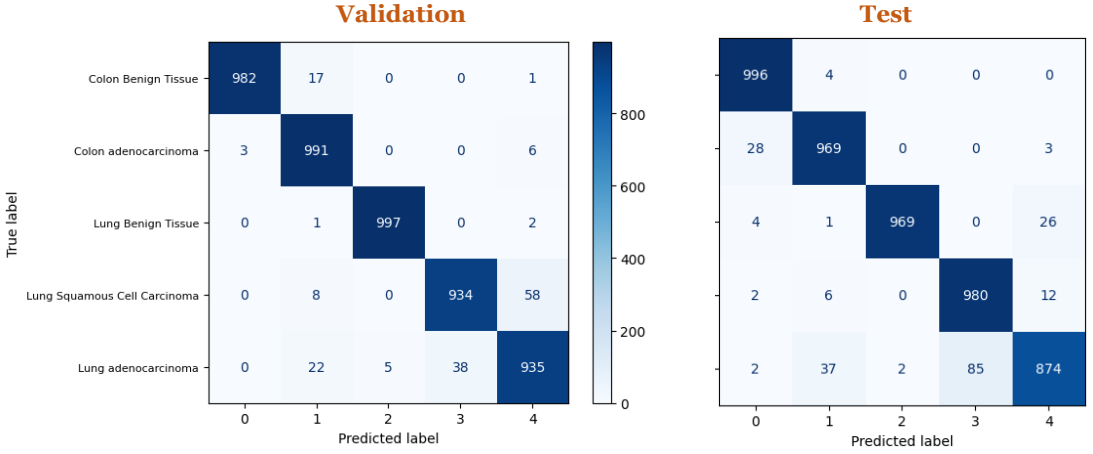


	Validation				Test			
	precision	recall	F1-score	support	precision	recall	F1-score	support
Colon benign tissue	0.98	1.00	0.99	1000	0.97	1.00	0.99	1000
Colon adenocarcinoma	1.00	0.98	0.99	1000	1.00	0.97	0.98	1000
Lung benign tissue	1.00	1.00	1.00	1000	1.00	1.00	1.00	1000
Lung squamous cell carcinoma	0.97	0.97	0.97	1000	0.95	0.98	0.97	1000
Lung adenocarcinoma	0.97	0.96	0.97	1000	0.98	0.95	0.96	1000
accuracy			0.98	5000			0.98	5000
macro avg	0.98	0.98	0.98	5000	0.98	0.98	0.98	5000
weighted avg	0.98	0.98	0.98	5000	0.98	0.98	0.98	5000

5. Transfer Learning Model and Its Results

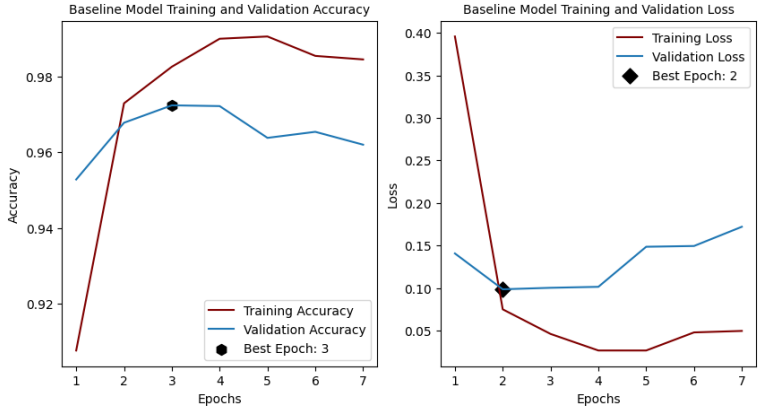
VGG16 model developed by the Visual Graphic Groups (Simonyan & Zisserman, 2014) was used for this step, original convolution and pooling layers of the VGG16 model was combined with the two dense layers with 128 and 64 units, along with the 5 units output layer with softmax activation function. Categorical cross entropy was selected as the loss function while the adam was used as the compiler optimizer. Model was trained using 32 epochs and 64 batch size while early stopping was introduced similar to the enhanced model.

Figure 9. Confusion matrix and performance measures for validation and test (transfer learning)



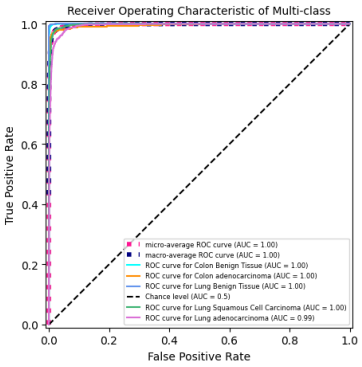
	Validation				Test			
	precision	recall	F1-score	support	precision	recall	F1-score	support
Colon benign tissue	1.00	0.98	0.99	1000	0.97	1.00	0.98	1000
Colon adenocarcinoma	0.95	0.99	0.97	1000	0.95	0.97	0.96	1000
Lung benign tissue	1.00	1.00	1.00	1000	1.00	0.97	0.98	1000
Lung squamous cell carcinoma	0.96	0.93	0.95	1000	0.92	0.98	0.95	1000
Lung adenocarcinoma	0.93	0.94	0.93	1000	0.96	0.87	0.91	1000
accuracy			0.97	5000			0.96	5000
macro avg	0.97	0.97	0.97	5000	0.96	0.96	0.96	5000
weighted avg	0.97	0.97	0.97	5000	0.96	0.96	0.96	5000

Figure 10. Transfer learning model loss and accuracy across epochs



Validation accuracy reached to 97% in the 3th epoch and started fluctuating afterwards. Although the training loss continued to improve as the number of epochs increased, the validation loss started to rise following the 2nd epoch, indicating an overfit (figure10). Overall, the enhanced model produced a more balanced performance to segment different classes (Figure 11) and reached to 96% average accuracy on the test data (Figure 9).

Figure 11. RoC curve for the test data (transfer learning)



6. Discussion

Average test accuracy has been improved from 72% at the baseline to the 98% using enhanced model. Besides the increase in the average accuracy, model performance on classifying images belonged to the different tissue groups was also enhanced (see figures 5 and 7). While the performance of the hyperparameters are dependent to the cross-combinations, increasing number of filters in the second convolution layer has substantially improved the model performance, likewise introducing the third convolution layer has also contribute to overall performance and the average validation accuracy exceeded the 95% threshold. Even though increasing the depth of CNN may result in an overfit and poor generalization capacity; introducing the dropout layers and using early stopping has extended our control over the model. Enhanced model was also slightly outperformed the transfer learning model with the 97% test accuracy on average. A further comparison across the classes, reveals that the model performance to classify images belonged to lung adenocarcinoma lags behind the other classes in both models while the performance gap is more evident for the transfer learning with 87% recall rate. On the other hand, transfer learning converged significantly faster compared to the enhanced model where minimum loss and the maximum accuracy were recorded at the 2nd and 3rd epochs, respectively.

To further improve the performance of the enhanced model, following strategies may be incorporated: (i) using RMSprop as optimizer (Masud et al., 2021; Masud et al., 2020; Mangal, Chaurasia & Khajanchi, 2020), (ii) implementing batch minimization, (iii) adding connector convolution blocks (Masud et al., 2020), (iv) pre-processing images with sharper masking (Masud et al., 2021).

7. References

Mangal, S., Chaurasia, A., & Khajanchi A. (2020). Convolutional Neural Networks for Diagnosing Colon and Lung Cancer Histopathological Images. arXiv preprint arXiv:2009.03878v1.

Masud, M., Sikder, N., Nahid, A.A., Bairagi, A.K., & Alzain, M.A. (2021). A Machine Learning Approach to Diagnosing Lung and Colon Cancer Using a Deep Learning-Based Classification Framework. *Sensors (Basel, Switzerland), 21*.

Masud, M., Muhammad, G., Hossain, S.M., Alhumyani, H., Sultan, S.A., Cheikhrouhou, O., Ibrahim, S. (2020). Light Deep Model for Pulmonary Nodule Detection from CT Scan Images for Mobile Devices. *Wireless Communications and Mobile Computing,* 2020.

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.