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:
Teresa Virca (2090933)

March 2024

1. Problem Definition

As machine learning techniques become more advanced, they can bring about important improvements in the field of medicine (Borkowski et al., 2019). In particular, machine learning can be useful for analyzing images and spotting patterns which might not be visible to the human eye. The image dataset LC25000, was created for the study of cancer pathology, and consist of images belonging to five classes: colon adenocarcinoma, benign colonic tissue, lung adenocarcinoma, lung squamous cell carcinoma, and benign lung tissue. The present study will attempt to correctly classify the different classes of tissues by implementing convolutional neural networks (CNNs). If successful, such a model can become a supportive tool for cancer pathologists.

2. Exploratory Data Analysis

The dataset contains 750 original images of lung tissue and 500 of colon tissue. Image augmentation was used to create a total dataset of 25.000 images (Borkowski et al., 2019). As part of pre-processing, the images were resized to 120x120. The five classes are equally distributed, with 5.000 images for each of the five classes present, as shown in Figure 1. To visualize the class differences, 15 random samples were plotted with the corresponding labels, as shown in Figure 2.

As part of the pre-processing, the classes were one-hot encoded. Class 0 corresponds to 'Colon Benign Tissue' , class 1 to 'Colon adenocarcinoma', class 2 to 'Lung Benign Tissue', class 3 to 'Lung Squamous Cell Carcinoma' and class 4 to 'Lung adenocarcinoma'. The data was then split into a training (60%), validation (20%) and test (20%) set, by means of stratified splits. The random state 42 was assigned for reproducibility. The training set contains 16.000 rows, the validation set 4.000 rows, and the test set 5.000 rows. Even though, normalization could be employed to rescale the pixel values, this step was omitted, as per instructions.

3. Results of the Baseline Model

The baseline model consists of two convolutional layers with ReLU activation followed by MaxPooling layers. The MaxPooling layers are followed by two dense layers with ReLU activation. Finally, an output layer of size 5, given the 5 output classes, with Softmax activation function. This activation function was chosen given that it is suitable for a multi-class classification problem, as the function produces the probability distribution over the classes, assigning the greatest probability to the correct class (Goodfellow et al., 2016). Moreover, padding was applied to the convolutional layers, so as to maintain the dimensionality unchanged. The model was compiled using the Adam optimizer, the loss function was specified as categorical cross-entropy, and accuracy was set as metric for evaluating the performance of the model. The model was then trained on 10 epochs and using a batch size of 32.

The accuracy on the validation set was registered as 0.71, while the accuracy on the test set was registered as 0.72. Figure 3 shows the training process throughout the epochs. In the lower image, the training accuracy continues increasing, whereas the validation accuracy is more unstable as it increases and decreases at various epochs. It is evident by looking at the validation loss and accuracy values that the model is overfitting, and should therefore be adjusted accordingly. Figure 4 shows the classification report for the validation set. The F1-score for the five classes ranges from 0.60 to 0.89, highlighting differences between the precision and recall of the various classes. Specifically, colon adenocarcinoma and lung adenocarcinoma report low values, which means that affected tissues are not classified correctly. The confusion matrix for the validation set in Figure 5 shows that colon adenocarcinoma was particularly often misclassified as colon begging tissue, and vice versa. This means that the model often classifies affected tissues as benign, which is particularly problematic, and vice versa. The matrix also shows that lung squamous cell carcinoma was often classified as lung adenocarcinoma, and vice versa. This means that the model incorrectly classifies different types of affected lung tissue. Finally, the Receiver Operating Characteristic (ROC) curves for the validation set shown in Figure 6 are leaning towards the upper left corner of the graph, indicating that the results are not random. The Area Under the Curve (AUC) values for the five classes range between 0.91 and 0.98, in particular the classification is more accurate for lung benign tissue, and least accurate for colon adenocarcinoma and lung adenocarcinoma. This indicates once again that the model does not classify tissues affected by cancer correctly.

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

The initial improvement to the baseline model aims to change the architecture and increase the depth of the network, drawing inspiration from the VGG16 model by Simonyan and Zisserman (2014). VGG16 uses a simple but very deep architecture, and small-sized filters which contribute to the receptivity of the model to patterns in the images. The implemented model comprises of 4 blocks of convolutional and MaxPooling layers, and the filter sizes increase gradually after each block (32, 64, 128, 256). Each convolutional layer employs a 3x3 filter with ReLU activation and padding is used to maintain spatial dimensions. Max-Pooling layers with a pool size of 2x2 and a stride of 2x2 are used to reduce spatial dimensions. Following the convolutional blocks, the data is flattened and directed through two fully connected layers, each with 128 neurons and ReLU activation. The final output layer consists of five neurons with softmax activation, catering to a multi-class classification problem. This architecture aims to capture more intricate hierarchical features in the images compared to the baseline model. This initial improvement yields an accuracy of 0.97 on both the validation and test set, signaling the importance of depth in convolutional networks for image recognition.

This architecture was further improved by reducing the number of convolutional layers in the first two blocks and reducing the number of neurons in the fully connected layers in order to reduce computational complexity. Decreasing the number of parameters serves to enhance the model's generalizability and avoid overfitting. It also leads to faster training times and potentially better convergence, as the model is less likely to memorize noise in the training data. The improved architecture, which is shown in Figure 7, consists of 4 convolutional and max-pooling blocks. The first two blocks include convolutional layers with 32 and 64 filters respectively, followed by max-pooling with a 2x2 window. Subsequent blocks follow a similar pattern but double the number of layers and increase the number of filters to 128, and 256, respectively.

L2 regularization with a coefficient of 0.001 is applied to the convolutional layers to enhance robustness. The literature suggests that setting a lower weight regularization parameter can help the model learn (Krizhevsky et al., 2012; Simonyan & Zisserman, 2014). The fully connected layers include two dense layers with 64 neurons each, incorporating dropout with a rate of 0.2. Adding dropout to fully connected layers introduces a form of stochastic regularization during training. It randomly drops out a fraction of neurons during each training iteration, preventing co-adaptation of neurons and reducing the risk of overfitting. This technique forces the model to learn more robust features and improves its ability to generalize (Goodfellow et al., 2016). Dropout rates between 10 and 30% were considered, however, the 20% rate proved to be the most effective. The output layer, utilizing the softmax activation function, comprises five neurons corresponding to the distinct classes in the classification task. The Adam optimizer is employed and the learning rate is set to 0.0001. Various learning rates were tested, starting from 0.1 and exponentially decreasing, as this is a common practice (Wu et al., 2019). Moreover, it was observed that higher learning rates cause the loss to decrease abruptly, while failing to make further improvements throughout the remaining epochs. Finally, the number of epochs is set to 25 and early stopping is implemented with a patience of 5 based on validation accuracy. When running the model multiple times, the model always stops before the last epoch, signaling that early stopping helps the model to end the training process when the validation accuracy does not improve for over 5 epochs.

The final fine-tuned model has a validation and test accuracy of 0.98, which is an improvement over the initially improved structure. The cross entropy loss and classification accuracy throughout the training is displayed in Figure 8. Figure 9 shows the classification report for the validation set. The F1-score ranges between 0.96 for lung adenocarcinoma and 1 for lung benign tissue. The model improved considerably in recognizing images of colon adenocarcinoma, as the precision for the tuned model is 1, and the recall and F1-score are both 0.99. The confusion matrix for the validation set in Figure 10 shows that most samples were classified correctly, as the colors on the left-to-right diagonal are visibly more intense than the rest. A few misclassification errors are still present, especially for what regards lung squamous cell carcinoma and lung adenocarcinoma. Despite this, the model has greatly improved compared to the baseline model. Finally, the ROC curves for the validation set are shown in Figure 11. All classification lines are neatly packed in the upper left corner of the graph, and the AUC for all classes is 1. Despite this, it is still visible that predictions for the fourth class, namely lung adenocarcinoma, are slightly less accurate, as the previous figures showed.

5. Transfer Learning Model and Its Results

The chosen transfer learning model makes use of the VGG16 convolutional network developed by Simonyan & Zisserman (2014). The model is loaded with pre-trained weights and its layers are frozen to prevent further training. The architecture includes a flattening layer applied to the output of the VGG16 model, followed by two fully connected layers with 128 and 64 neurons, respectively, and ReLU activation functions. The final output layer consists of 5 neurons with a softmax activation. During training, the model uses the Adam optimizer and categorical cross-entropy loss function. An early stopping callback is employed to monitor validation accuracy, and the model is trained for 30 epochs with a batch size of 32. This approach leverages the pre-trained features of VGG16 for feature extraction and introduces task-specific learning through the subsequent dense layers.

The transfer learning model achieved an accuracy of 0.97 on both validation and test set, only slightly lower than the fine-tuned model. The cross-entropy loss and accuracy plot in Figure 12 highlights the presence of overfitting, as the training data achieves a perfect accuracy of 1, while the validation data significantly underperforms the training data. Figure 13 shows the classification report for the validation set. The F1-scores for the different classes ranges between 0.94 for lung adenocarcinoma and 0.99 for both lung and colon benign tissue. The confusion matrix for the validation set in Figure 15 shows that, similarly to the fine-tuned model, the transfer learning model mostly misclassifies samples of lung squamous cell carcinoma and lung adenocarcinoma. Finally, the ROC curve for the validation set in Figure 14, displays promising results, as the AUC value for all classes is 1. It is once again visible that lung adenocarcinoma has a slightly lower accuracy than the other classes.

6. Discussion

When training the models presented above, the greatest challenge was that of preventing overfitting. If the network learns too closely the patterns and idiosyncrasies of the training data, unseen samples are misclassified. The transfer learning model only slightly underperforms the improved model, which suggests that by further fine-tuning hyperparameters, better performance metrics could be obtained.

Even though the results presented here show promising results, further improvements should be considered for the classification problem at hand. Wang et al. (2017) propose a Residual Attention Network which introduces an attention mechanism within the convolutional network, which utilizes attention residual learning for training the network. Such a method could focus on relevant regions of the image and improve the diagnostic accuracy of the results and reduce the fraction of false negatives. Another approach could consider data augmentation for the generation of larger training sets. For example, Wolterink et al. (2018) introduce Generative Adversarial Networks for generating synthetic medical images, which can further diversify the training set thus producing more robust results. Finally, Gibson et al. (2017) point to the fact that many deep learning elements of medical image analysis entail domain-specific peculiarities. Because of this, a tailored network architectures could be developed for cancer pathology.

Figures

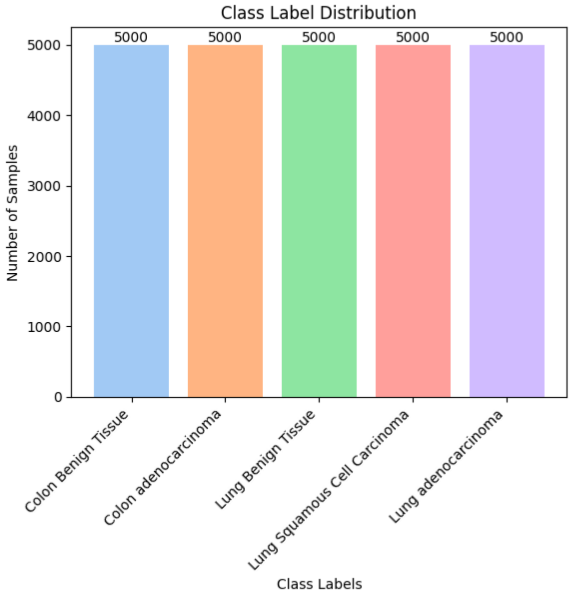


Figure 1: Distribution of Classes

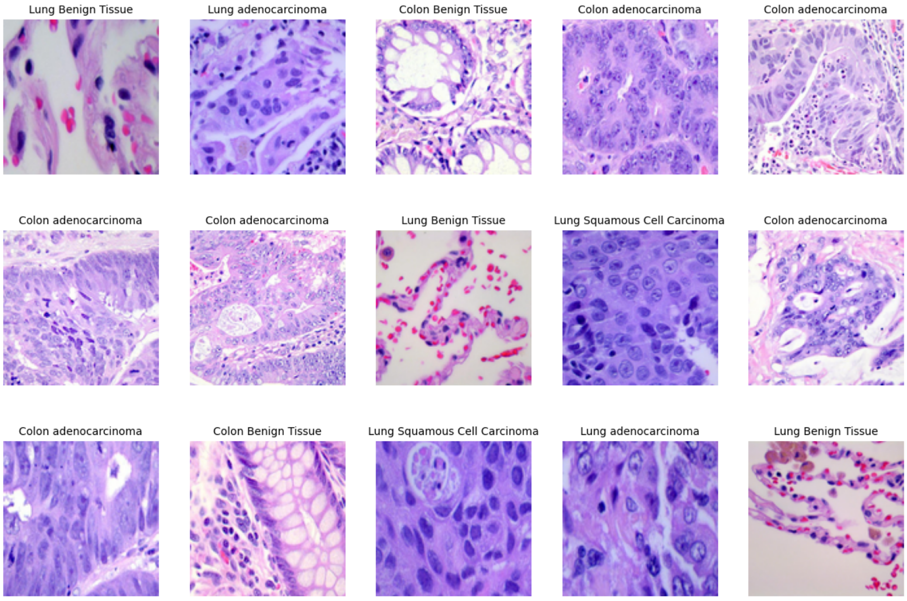


Figure 2: 15 Random Samples

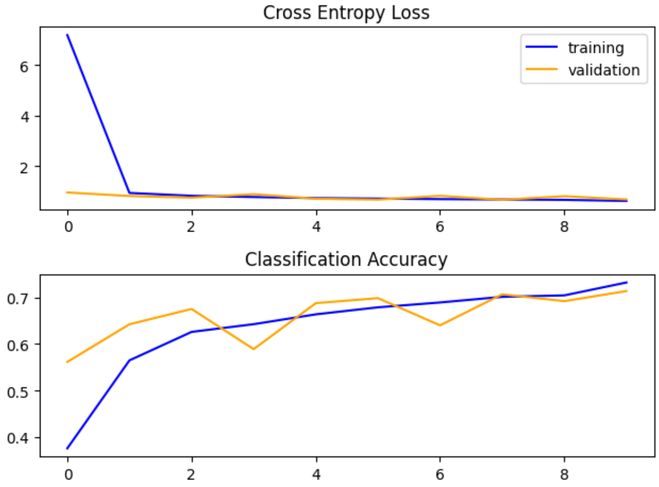
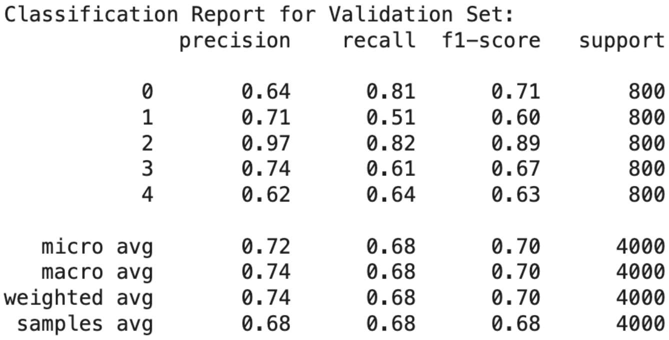


Figure 3: Baseline - Loss-Accuracy Plot



```
Classification Report for Validation Set:
              precision    recall  f1-score   support

           0       0.64      0.81      0.71       800
           1       0.71      0.51      0.60       800
           2       0.97      0.82      0.89       800
           3       0.74      0.61      0.67       800
           4       0.62      0.64      0.63       800

   micro avg       0.72      0.68      0.70      4000
   macro avg       0.74      0.68      0.70      4000
weighted avg       0.74      0.68      0.70      4000
 samples avg       0.68      0.68      0.68      4000
```

Figure 4: Baseline - Classification Report (Validation)

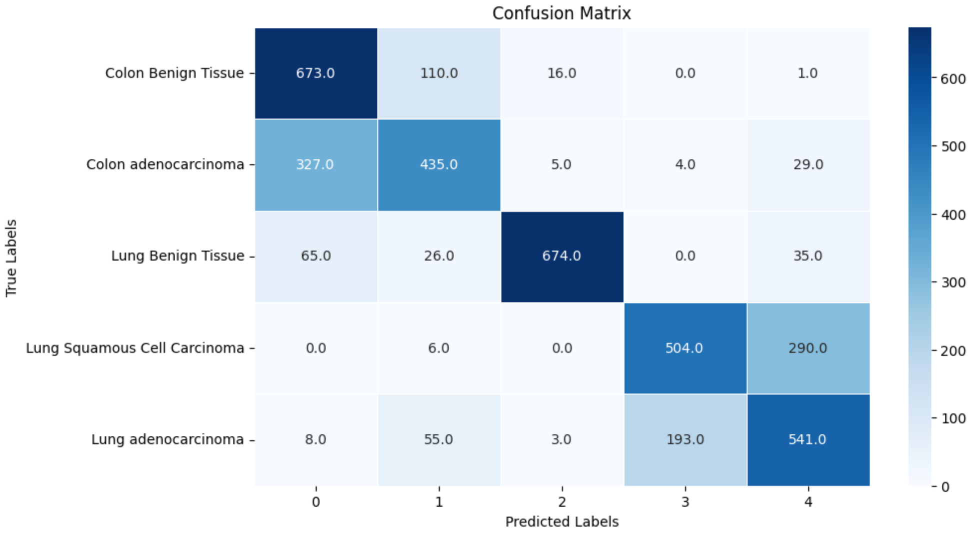


Figure 5: Baseline - Confusion Matrix (Validation)

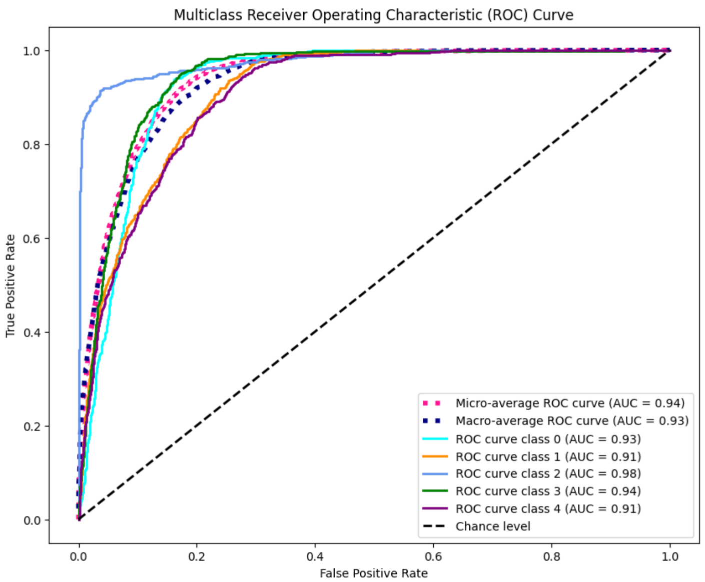
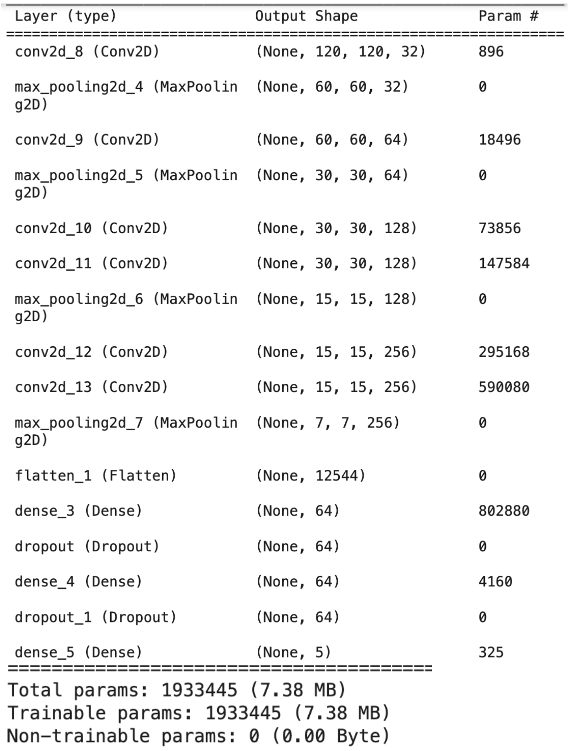
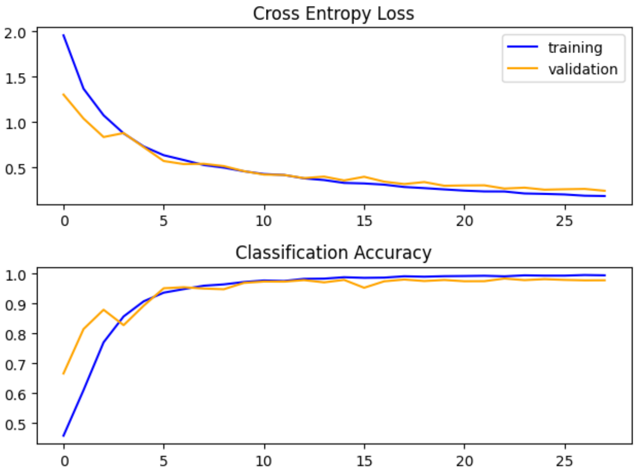


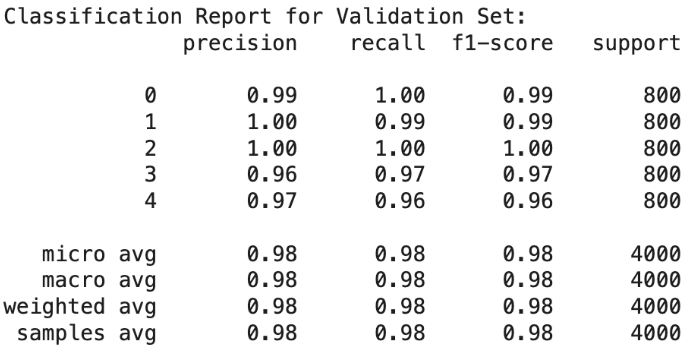
Figure 6: Baseline - ROC (Validation)



```
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_8 (Conv2D)           (None, 120, 120, 32)      896

 max_pooling2d_4 (MaxPoolin  (None, 60, 60, 32)        0
 g2D)

 conv2d_9 (Conv2D)           (None, 60, 60, 64)        18496

 max_pooling2d_5 (MaxPoolin  (None, 30, 30, 64)        0
 g2D)

 conv2d_10 (Conv2D)          (None, 30, 30, 128)       73856

 conv2d_11 (Conv2D)          (None, 30, 30, 128)       147584

 max_pooling2d_6 (MaxPoolin  (None, 15, 15, 128)       0
 g2D)

 conv2d_12 (Conv2D)          (None, 15, 15, 256)       295168

 conv2d_13 (Conv2D)          (None, 15, 15, 256)       590080

 max_pooling2d_7 (MaxPoolin  (None, 7, 7, 256)         0
 g2D)

 flatten_1 (Flatten)         (None, 12544)             0

 dense_3 (Dense)             (None, 64)                802880

 dropout (Dropout)           (None, 64)                0

 dense_4 (Dense)             (None, 64)                4160

 dropout_1 (Dropout)         (None, 64)                0

 dense_5 (Dense)             (None, 5)                 325
==================================
Total params: 1933445 (7.38 MB)
Trainable params: 1933445 (7.38 MB)
Non-trainable params: 0 (0.00 Byte)
```

Figure 7: Tuned Model - Summary

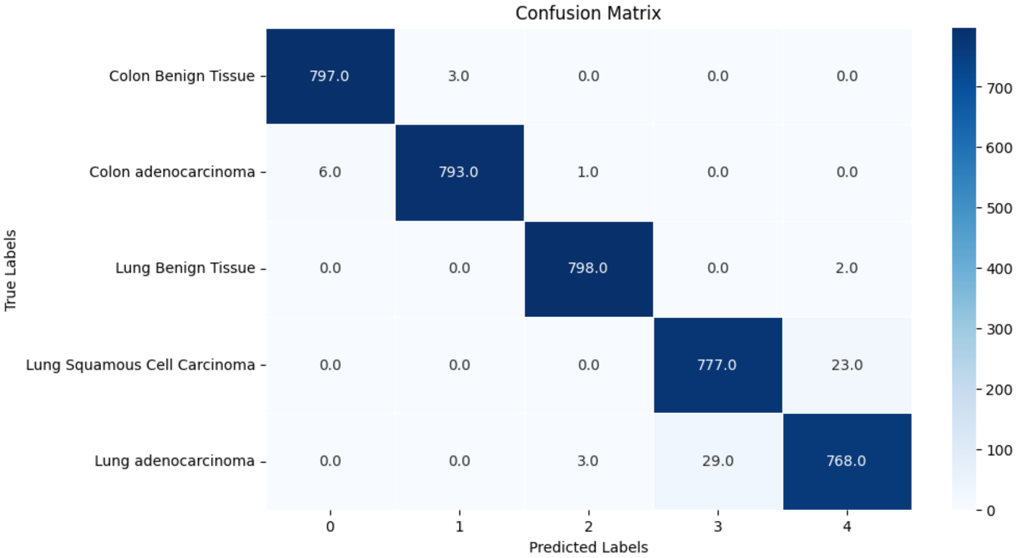




```
Classification Report for Validation Set:
             precision    recall  f1-score   support

           0       0.99      1.00      0.99       800
           1       1.00      0.99      0.99       800
           2       1.00      1.00      1.00       800
           3       0.96      0.97      0.97       800
           4       0.97      0.96      0.96       800

   micro avg       0.98      0.98      0.98      4000
   macro avg       0.98      0.98      0.98      4000
weighted avg       0.98      0.98      0.98      4000
 samples avg       0.98      0.98      0.98      4000
```

Figure 8: Tune Model - Loss-Accuracy Plot

Figure 9: Tuned Model - Classification Report (Validation)



	0	1	2	3	4
Colon Benign Tissue	797.0	3.0	0.0	0.0	0.0
Colon adenocarcinoma	6.0	793.0	1.0	0.0	0.0
Lung Benign Tissue	0.0	0.0	798.0	0.0	2.0
Lung Squamous Cell Carcinoma	0.0	0.0	0.0	777.0	23.0
Lung adenocarcinoma	0.0	0.0	3.0	29.0	768.0

Figure 10: Tuned Model - Confusion Matrix (Validation)

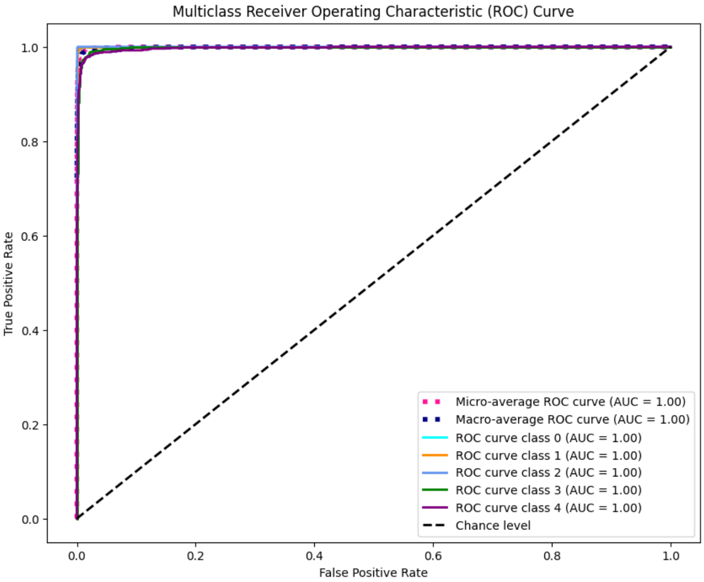


Figure 11: Tuned Model - ROC (Validation)

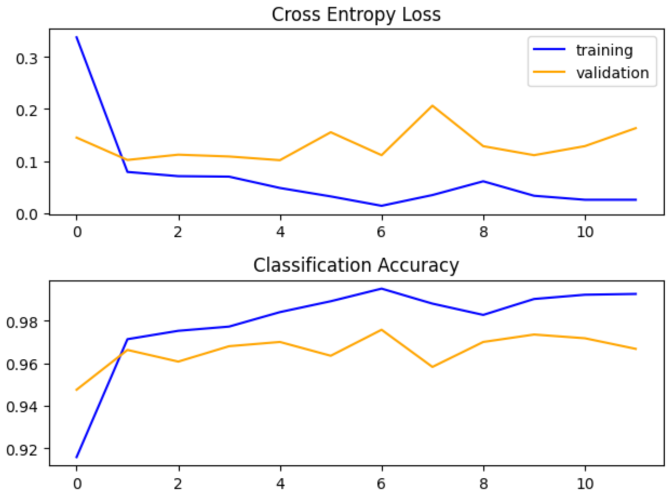
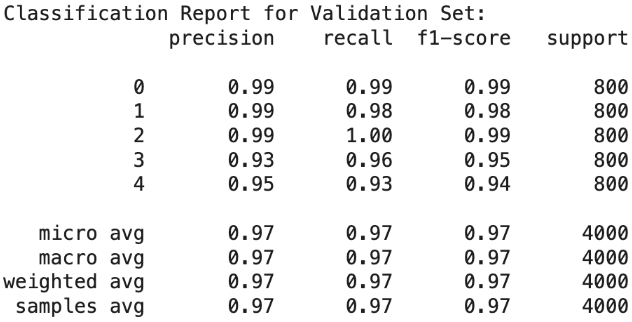


Figure 12: Transfer Learning - Loss-Accuracy Plot



```
Classification Report for Validation Set:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       800
           1       0.99      0.98      0.98       800
           2       0.99      1.00      0.99       800
           3       0.93      0.96      0.95       800
           4       0.95      0.93      0.94       800

   micro avg       0.97      0.97      0.97      4000
   macro avg       0.97      0.97      0.97      4000
weighted avg       0.97      0.97      0.97      4000
 samples avg       0.97      0.97      0.97      4000
```

Figure 13: Transfer Learning - Classification Report

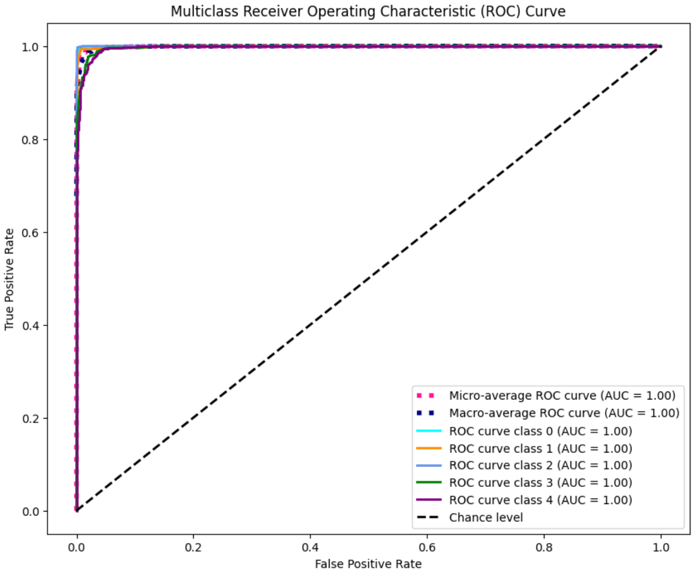
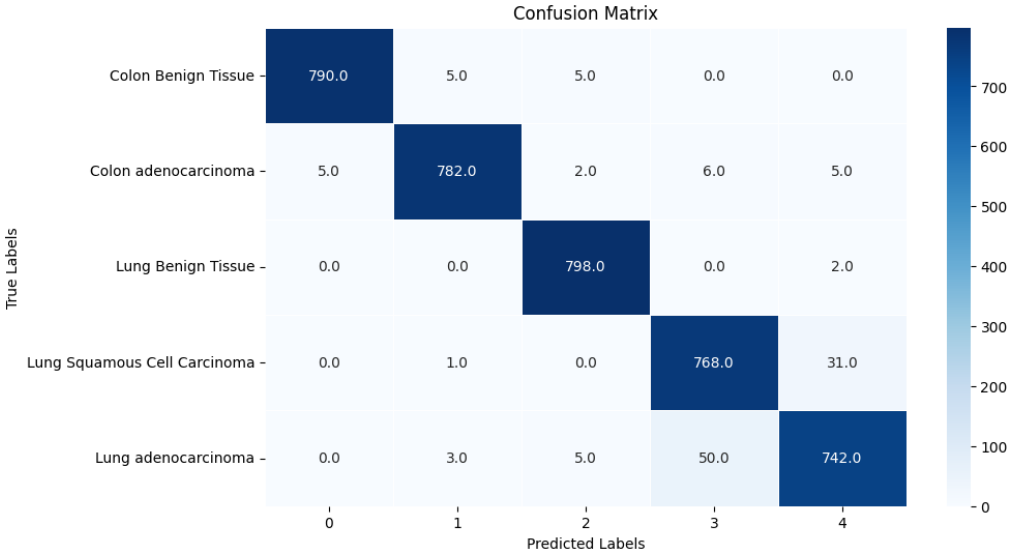


Figure 14: Transfer Learning - ROC (Validation)



	0	1	2	3	4
Colon Benign Tissue	790.0	5.0	5.0	0.0	0.0
Colon adenocarcinoma	5.0	782.0	2.0	6.0	5.0
Lung Benign Tissue	0.0	0.0	798.0	0.0	2.0
Lung Squamous Cell Carcinoma	0.0	1.0	0.0	768.0	31.0
Lung adenocarcinoma	0.0	3.0	5.0	50.0	742.0

Figure 15: Transfer Learning - Confusion Matrix (Validation)

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 preprint arXiv:1912.12142*.

Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D. I., Wang, G., … & Vercauteren, T. (2018). NiftyNet: a deep-learning platform for medical imaging. *Computer methods and programs in biomedicine*, *158*, 113-122.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems. 25*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. a*rXiv preprint arXiv:1409.1556*.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., ... & Tang, X. (2017). Residual attention network for image classification. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156-3164).

Wolterink, J. M., Kamnitsas, K., Ledig, C., & Išgum, I. (2018). Generative adversarial networks and adversarial methods in biomedical image analysis. *arXiv preprint arXiv:1810.10352*.

Wu, Y., Liu, L., Bae, J., Chow, K. H., Iyengar, A., Pu, C., Wenqui, W., Lei, Y. & Zhang, Q. (2019, December). Demystifying learning rate policies for high accuracy training of deep neural networks. *In 2019 IEEE International conference on big data (Big Data),* (pp. 1971-1980). IEEE.