COSC 2673/2793 | MACHINE LEARNING

Assignment 2: Classify Images of Cancer

Written by:

Aayushi Khatri: s3948240

Sarang Kuniyil: s3914175

Introduction

Histopathology is the study of tissue and cells under a microscope to diagnose diseases. Recently, advances in digital imaging technology have made it possible to use machine learning algorithms to automate the analysis of histopathology images. This report attempts to develop a machine learning model that can classify histopathology images of cancerous cells using a modified version of the publicly available *CRCHistoPhenotypes dataset*. Our objective is to create a robust machine learning model that can accurately classify histopathology images of cancerous cells and classify images according to the cell type, which can have far-reaching implications in cancer diagnosis and treatment.

Findings of Exploratory Data Analysis (EDA)

Based on the exploratory data analysis conducted, we found that there are 17.56% more non-cancerous images than there are cancerous images (appendix A). Additionally, the number of samples for different cell types is not equal, with epithelial cells having the highest number of impressions (appendix B). The unbalanced impressions in each of these classes will inform our choice of evaluation metrics for the model. It is also apparent that all epithelial cells in the dataset are cancerous (appendix C), although real-world research indicates that this is not a reliable rule of thumb as everyone has epithelial cells which are not always malignant. This is a potential pitfall of the data as the model might associate epithelial cells with cancer, which could lead to problems when deploying the model in the real world.

Choice of Model

The Convolutional Neural Network (CNN) deep learning model is a highly effective approach for image classification tasks. CNN models are specifically designed to analyse and extract features from images and can learn patterns and relationships between pixels and features at different levels of abstraction. CNN models are also highly scalable, allowing them to analyse large datasets and learn complex patterns that would be difficult to detect using traditional machine learning algorithms. While CNN models can be expensive to train as they require more computational power, that is a trade-off we are willing to make as CNN models have achieved good performance on various image classification tasks, including histopathology image classification, in other published research papers. A visualisation of the final CNN model architecture for cancer classification and cell type classification is included in appendix D and E, respectively. Decision trees were briefly considered for the cell type classification model but were not suitable due to its lack of robustness. A study conducted by Statnikov et. al (2008) found that decision trees underperformed on image classification tasks, therefore we did not attempt to use it.

Evaluation Framework

The evaluation metrics we will use to assess the success of the model are:

Loss - is a fundamental metric that represents the discrepancy between the predicted output and the true output. In classification tasks, the loss is typically computed using a specific loss function. For the cancer classification model, we will be using the binary cross-entropy loss function as it is appropriate for binary classification and for the cell type classification model, we will use the adam loss function as it most appropriate for models classifying more than two classes. Our goal is to minimize the loss, as lower loss values indicate better alignment between the predicted and true outputs.

F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a single metric that combines both precision and recall, offering a balanced assessment of the model's performance. The F1 Score is particularly useful when the number of impressions in classes are imbalanced, as seen in through the EDA.

For the Cancer Classification model, we expect to see an F1 Score of 85% and a loss of 30%. For the Cell Classification model, we expect to see an F1 Score of 80%, and a loss of 50%.

Our Approach and Iterations

Cancer Classification Model

We utilised the TensorFlow image dataset API through the Keras utility module. This took the classified data from two separate directories (cancerous and non-cancerous classes) and converted them into NumPy arrays that could be iterated through in batches of size 32. We did not change the default batch size as it is commonly used for image classification tasks and would also give the model more diverse samples to learn from. We then scaled all the NumPy arrays by dividing the value by 255, to scale the data between 0 and 1. This was mainly a time-saving measure as the CNN would run faster with smaller values, however it also meant the model would not incorrectly value one learnable parameter over the other.

There was a total of 310 batches in our data set (images that we had labels for), and we chose to split them into 3 data sets, following the 60-20-20% approach, where 60% of data was moved to the training set, 20% was for validation during training and 20% was kept aside for testing. The baseline model included 3 convolution layers that captured local patterns and

features in the input images, followed by a max pooling layer to reduce the spatial dimensions of the feature maps and extract the most relevant information. We also incorporated a fully connected dense layer, which allowed the model to capture global relationships within the data. The flattening layer transformed the 2D feature maps into a 1D vector which made them suitable to feed into a dense layer. Lastly, we used the ReLU activation functions for all the convolution layers as it would help introduce non-linearity to the model. However, we changed the activation function for the last layer to Sigmoid because we thought it was the most appropriate for a binary classification problem like this one. The loss function we used to compile the model was the binary cross-entropy function as it was the most appropriate for a binary classification task. As we trained the baseline model, for 20 epochs, we noticed the model was overfitting (appendix F). We attributed this problem to the model running for too many epochs and there not being any form of regularisation. To The updated model (architecture shown in appendix D) only differs from the baseline model in the way that there is a dropout layer that is added before the output layer, which reduces the complexity of the model. We also used an early stopping call back to fine tune the number of epochs, monitoring the validation dataset's loss. After using these techniques, the updated model performed significantly better (appendix G) The gap between the training metrics and validation metrics decreased, indicating that the model was not overfitting as much as before.

Cell Type Classification Model

For the cell type classification model, we followed the same data pre-processing, splitting, and scaling approach. However, the building of the model architecture was different. The baseline model made use of VGG blocks. One VGG block in the model consists of two convolution layers, which learned, and extracted local features form the input image, and a max pooling layer, which reduced the spatial dimensions of the feature maps, capturing the most important features while discarding some spatial information. The convolution layers are activated with the ReLU function. Another advantage of using a ReLU activation function is that due to its non-saturating nature it allows for easier learning in deep architectures. The gradients can flow more easily through the layers, facilitating the training of deep CNN models like this one. The last VGG block in the baseline model has a flattening layer that feeds into the dense layer. Lastly, this block predicts whether the image falls into one of the four classes. The baseline model ran for 25 epochs initially and we noticed there was overfitting (appendix H). Therefore, we attempted to reduce overfitting by using early call backs that stopped the model after the validation loss did not improve after 3 consecutive epochs (the patience value). This led to a much lower validation loss value of 0.66, compared to the 1.655 validation loss of the baseline model. For the next iterations of the model, we attempted to regularise it in two difference ways. Firstly, we used L2 regularisation with every convolution layer as it would add penalty term to the loss function, encouraging the model to use smaller weights. This would lead to a simpler model that would tune out the noise or irrelevant features. Secondly, we added a dropout layer before the final output layer to further simplify the model. Following these improvements, overfitting was reduced however the evaluation metrics did not improve.

Evaluation of Models and Ultimate Judgement

After finetuning the baseline model and adding a regularisation measure, we tested the cancer classification model. The final model met the evaluation framework set previously. The test loss was 0.3017, which is the lowest we have seen so far, and it meets the prediction of the loss being 0.30. This indicates that the model's ability to minimise the discrepancy between the predicted and actual values. The test F1 score is 0.9056 which is more than the expected 0.85 indicating that the model effectively predicts whether an image is cancerous or not. Based on these findings, it can be concluded that the model has performed well and achieved a high level of accuracy in classifying cancerous and non-cancerous cells. It demonstrates a strong predictive capability and meets the expected performance thresholds outlined in the framework and is therefore the ultimate judgement for the cancer classification task.

The second model classified the images based on cell type. The final model produced a validation F1 score of 0.6257, which was lower than expected, by 0.18. This indicates that the model has a moderate balance between precision and recall. The test loss was also higher than the expected at 0.7653. We believe this is partly due to the imbalanced samples in each cell type class as we did not take any steps to combat that issue with the data. Although the evaluation results are not an improvement from just running the baseline model with early stopping, the gap between the training and validation metrics is considerably less (appendix I), indicating that the effects of overfitting have reduced compared to the baseline model. Thus, we consider the final model, with regularisation in each layer and a dropout layer before the output layer, to be the best model out of all the iterations. Although the finetuning of the model did lower the validation loss from the baseline model without early stopping, the final model (appendix E) still underperforms. However, we consider it to be the best model that we could make as further optimisation was not feasible due to time constraints.

Analysis and comparison between classifying two categories (isCancerous and cellType Classification)

The cancerous classification model has a high F1 score of 0.9056 and a comparatively low loss score. The high F1 score means that the model has performed well at determining whether an image represents a malignant cell or not, while the low loss indicates that the model's predictions are quite near to the true values. Overall, the model performs well when it comes to correctly categorising cancerous cells.

In comparison to the cancer classification model, the cell type classification model has a larger loss score. Contrary to the cancer classification model, this model has lower precision, recall, and F1 scores. The model accurately detects the positive class (cell type) approximately 65% of the time, while the recall of 0.8167 suggests that the model captures about 81% of the positive class instances. The F1 score, which combines precision and recall, is 0.6623. Overall, the cell type classification model has lower performance metrics compared to the cancerous classification model. It suggests that the model may face challenges in accurately classifying cell types, potentially due to the complexity and variability of cell types within the dataset.

In conclusion, the cancerous classification model outperforms the cell type classification model in terms of both validation loss and F1 score. This indicates that the cancerous classification model is more reliable and effective in predicting whether an image represents a cancerous cell or not.

Independent Evaluation

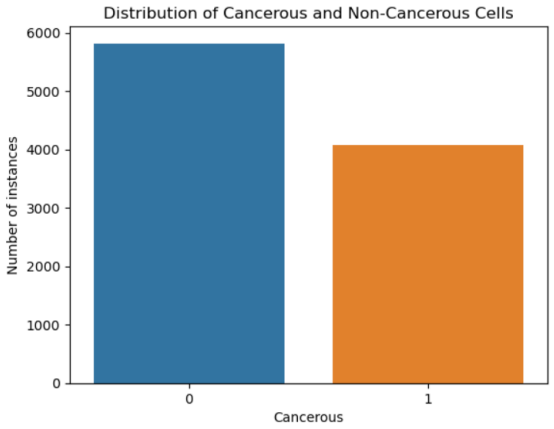
The original data we used for the exploration of this task was the publicly available *CRCHistoPhenotypes dataset.* Since the original publication of the research paper, many other published research papers have used this dataset. We will now compare the results of some of these papers with our findings.

In a paper titled, “Impact of Fully Connected Layers on Performance of Convolutional Neural Networks for Image Classification”, the authors’ objective was to investigate the impact of fully connected (FC) layers on the performance of CNNs for image classification. The authors varied the CNN model architecture by changing the number and positions of fully connected layers. The used four widely used datasets including, CIFAR-10, CIFAR-100, Tiny ImageNet, and CRCHistoPhenotypes, employing classification accuracy as their evaluation metric. Our evaluation metrics differ because our objective was to perform cancerous and cell type classification on imbalanced classes. Since, CRCHistoPhenotypes dataset, was the only one we used we needed to use an evaluation metric that accounted for the imbalanced classes, consequently we employed loss and F1 scores. The study concluded that for better performance, shallow CNNs require more nodes in FC layers, while deeper CNNs require fewer nodes. Deeper CNNs performed better than shallow models on deeper datasets, whereas shallow architectures performed better on wider datasets. The implication of this finding for new CNN models is the importance of finding the right balance between model complexity and overfitting to achieve better accuracy in CNN image classification.

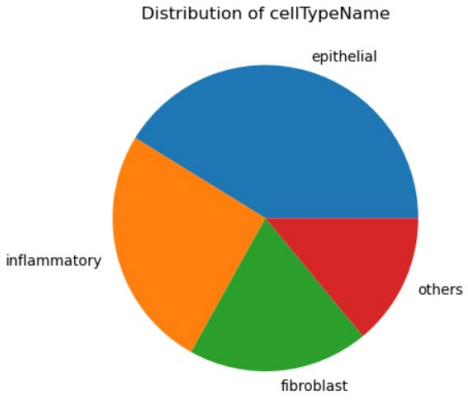
Another research paper that used the same dataset was the “RCCNet: An Efficient Convolutional Neural Network for Histological Routine Colon Cancer Nuclei Classification”. The paper’s objective was to classify histological cell nuclei reliably with a simple CNN model. Subsequently, their proposed RCCNet model consisted only of 1,5145,868 learnable parameters. They evaluated their model using a weighted average F1 score and classification accuracy, achieving a classification accuracy of 80.61% and a weighted F1 score of 0.7887. Their results demonstrated that RCCNet outperformed existing approaches in histological routine colon cancer nuclei classification. In comparison, our model underperformed based on their evaluation metrics.Their model differed from ours as it incorporated both local and global context information to predict the classification of the image. Additionally, their data pre-processing methodology included data augmentation techniques, transforming existing images in the data set to artificially increase the size of the dataset. We did not attempt data augmentation due to its advanced nature, which was beyond our capabilities. However, we recognise that it could have been beneficial to our models, both cancerous and cell type classification, as data augmentation could have helped rectify the imbalanced class distribution and addressed the issue of overfitting better. Ultimately, the RCCNet model demonstrated high classification accuracy and F1 score, indicating that the model’s precision and recall was balanced, surpassing the performance of previous methods.

Lastly, we examined the paper titled “Multi-level Feature Fusion for Nucleus Detection in Histology Images Using Correlation Filters”. The authors leverage multiple CNN architectures to extract features at different scales. They propose a fusion technique that combines these features to enhance nucleus detection accuracy. Their chosen evaluation metrics are precision, recall and F1 Score. These are similar to our evaluation metrics, as we also extracted precision and recall in order to calculate the F1 Score. The paper concludes that the multi-level feature fusion approach using correlation filters is effective for nucleus detection in histology images. The fusion of features at different scales enhances the model's ability to accurately identify and classify nuclei, providing valuable support for histological analysis and diagnosis.

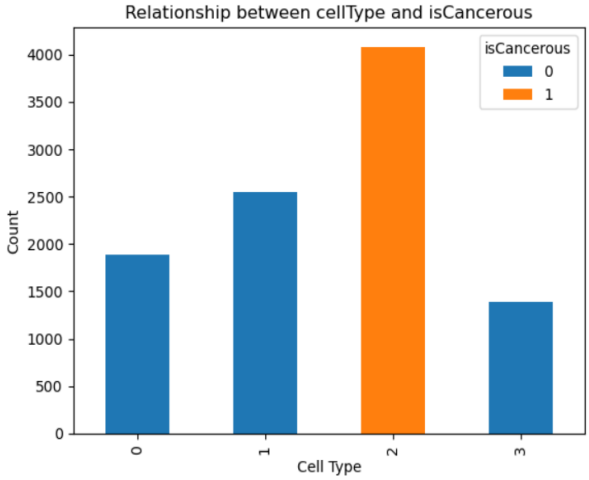
Appendix



Appendix A: Distribution of Cancerous and Non-cancerous cells



Appendix B: Distribution of cell types



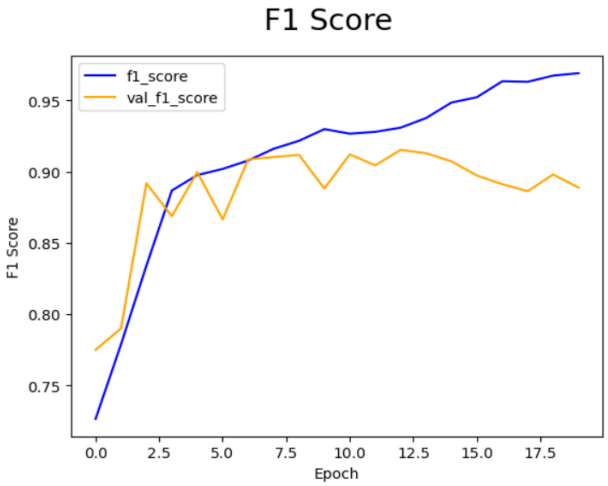
Appendix C: Relationship between cellType and isCancerous

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_3 (Conv2D)           (None, 254, 254, 16)      448

 max_pooling2d_3 (MaxPooling  (None, 127, 127, 16)     0
 2D)

 conv2d_4 (Conv2D)           (None, 125, 125, 32)      4640

 max_pooling2d_4 (MaxPooling  (None, 62, 62, 32)       0
 2D)

 conv2d_5 (Conv2D)           (None, 60, 60, 16)        4624

 max_pooling2d_5 (MaxPooling  (None, 30, 30, 16)       0
 2D)

 flatten_1 (Flatten)         (None, 14400)             0

 dense_2 (Dense)             (None, 256)               3686656

 dropout (Dropout)           (None, 256)               0

 dense_3 (Dense)             (None, 1)                 257

=================================================================
Total params: 3,696,625
Trainable params: 3,696,625
Non-trainable params: 0
```

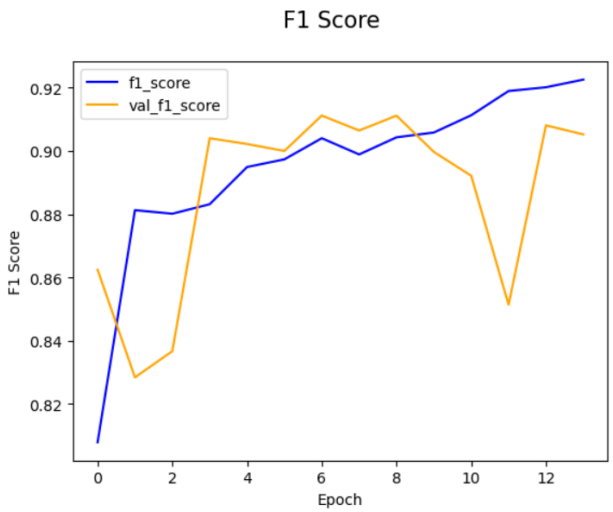
Appendix D: Cancer classification model architecture

```
Model: "sequential_23"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_138 (Conv2D)         (None, 32, 32, 32)        896

 conv2d_139 (Conv2D)         (None, 32, 32, 32)        9248

 max_pooling2d_69 (MaxPoolin  (None, 16, 16, 32)       0
 g2D)

 conv2d_140 (Conv2D)         (None, 16, 16, 64)        18496

 conv2d_141 (Conv2D)         (None, 16, 16, 64)        36928

 max_pooling2d_70 (MaxPoolin  (None, 8, 8, 64)         0
 g2D)

 conv2d_142 (Conv2D)         (None, 8, 8, 128)         73856

 conv2d_143 (Conv2D)         (None, 8, 8, 128)         147584

 max_pooling2d_71 (MaxPoolin  (None, 4, 4, 128)        0
 g2D)

 flatten_23 (Flatten)        (None, 2048)              0

 dense_46 (Dense)            (None, 128)               262272

 dropout_6 (Dropout)         (None, 128)               0

 dense_47 (Dense)            (None, 4)                 516

=================================================================
Total params: 549,796
Trainable params: 549,796
Non-trainable params: 0
_________________________________________________________________
```

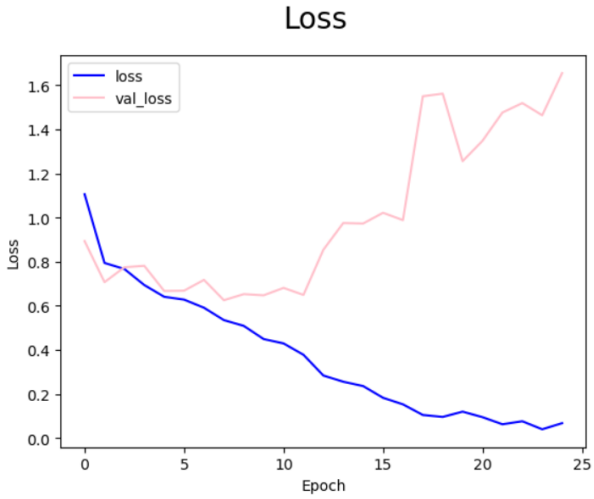
Appendix E: Cell type classification model architecture



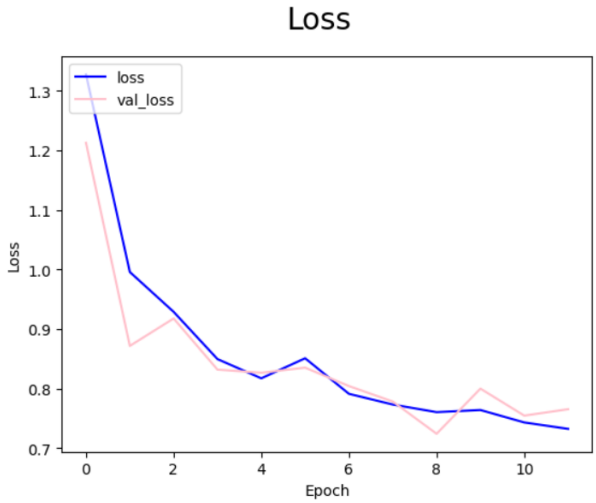
Appendix F: F1 Baseline Cancer Classification Model



Appendix G: F1 Score Final Cancer Classification Model Model



Appendix H: Loss Baseline Cell Type Classification



Appendix I: Loss Final Cell Type Classification Model

References

Alavi, A. (n.d.). *Bitbucket*. [online] bitbucket.org. Available at: https://bitbucket.org/alavi_a/rmit_aalavi-cosc2673_2793/src/main/labs/week09_DeepLearning/ [Accessed 8 May 2023].

Alavi, A. (n.d.). *Bitbucket*. [online] bitbucket.org. Available at: https://bitbucket.org/alavi_a/rmit_aalavi-cosc2673_2793/src/main/labs/week10_RuleLearning/ [Accessed 8 May 2023].

Basha, S.H.S., Dubey, S.R., Pulabaigari, V. and Mukherjee, S. (2019). Impact of fully connected layers on performance of convolutional neural networks for image classification. *Neurocomputing*, 378. doi:https://doi.org/10.1016/j.neucom.2019.10.008.

Javed, S., Mahmood, A., Dias, J. and Werghi, N. (2022). Multi-level feature fusion for nucleus detection in histology images using correlation filters. *Computers in Biology and Medicine*, [online] 143, p.105281. doi:https://doi.org/10.1016/j.compbiomed.2022.105281.

Jung, H., Lodhi, B. and Kang, J. (2019). An automatic nuclei segmentation method based on deep convolutional neural networks for histopathology images. *BMC Biomedical Engineering*, [online] 1(1). doi:https://doi.org/10.1186/s42490-019-0026-8.

Li, S., Jiang, H. and Pang, W. (2017). Joint multiple fully connected convolutional neural network with extreme learning machine for hepatocellular carcinoma nuclei grading. *Computers in Biology and Medicine*, 84, pp.156–167. doi:https://doi.org/10.1016/j.compbiomed.2017.03.017.

Renotte, N. (2022). *nicknochnack/ImageClassification*. [online] GitHub. Available at: https://github.com/nicknochnack/ImageClassification.

Shabbeer Basha, S.H., Ghosh, S., Kishan Babu, K., Ram Dubey, S., Pulabaigari, V. and Mukherjee, S. (2018). RCCNet: An Efficient Convolutional Neural Network for Histological Routine Colon Cancer Nuclei Classification. *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. doi:https://doi.org/10.1109/icarcv.2018.8581147.

Sirinukunwattana, K., Raza, S.E.A., Tsang, Y.-W., Snead, D.R.J., Cree, I.A. and Rajpoot, N.M. (2016). Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. *IEEE Transactions on Medical Imaging*, 35(5), pp.1196–1206. doi:https://doi.org/10.1109/tmi.2016.2525803.

Statnikov, A., Wang, L. and Aliferis, C.F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, 9(1), p.319. doi:https://doi.org/10.1186/1471-2105-9-319.

van der Laak, J., Litjens, G. and Ciompi, F. (2021). Deep learning in histopathology: the path to the clinic. *Nature Medicine*, 27(5), pp.775–784. doi:https://doi.org/10.1038/s41591-021-01343-4.

Wahab, N., Khan, A. and Lee, Y.S. (2017). Two-phase deep convolutional neural network for reducing class skewness in histopathological images based breast cancer detection. *Computers in Biology and Medicine*, 85, pp.86–97. doi:https://doi.org/10.1016/j.compbiomed.2017.04.012.