Title: **Machine Learning Assignment 2**

Student Name and email (contact info):

Denster Joseph Frank

s3894695

s3894695@student.rmit.edu.au

Jihun Lee

s3753624

s3753624@rmit.edu.vn

Affiliations: RMIT University.

Date of Report: 16/05/2023

We certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in my submission. I will show I agree to this honor code by typing "Yes": *Yes*.

Research Goal

The research goal for this project is to develop and compare machine learning models for classifying histopathology images of cancerous cells, specifically for predicting whether a given cell image represents cancerous cells or not and classifying the cell-type. Additionally, the project aims to evaluate and compare the performance of the developed models with other published works that have used the same dataset.

Data EDA and Pre-Processing

To ensure a sufficient amount of data for training and testing, we split the mainData.csv into train_data, val_data, and test_data using a **3:1:1** ratio. This ratio provides a large enough dataset for training, while also having ample data for testing the model's performance on unseen data. After splitting the data, we conducted EDA using the train_data to avoid data leakage. Using the plt.imread function, we confirmed that each image in the dataset had a size of 27x27 with 3 RGB colors **(27, 27, 3)**.

For evaluation, we first graphed the loss and accuracy of the model to determine whether it was overfitted or not. We then used the .evaluate function to check the final accuracy of the model using the test_data. We used to see if **F1 score** was improving from previous or other models. These evaluation metrics can help us determine the performance of the model and compare it with other models that have been tested on the same dataset.

Task 1: Classifying Images into Cancerous or isCancerous cell

A **baseline** binary classification model was created as a standard model for comparison with other models. For optimization, the model was compiled with the Adam optimizer, known for its adaptive learning rate. A learning rate of 0.0001 was set to control parameter updates By learning from the data over time and capturing complex patterns, the model was trained for 25 epochs. It achieved an accuracy rate of 87.4%, which served as an initial benchmark for comparison and further improvement. It includes a dense output layer with sigmoid activation, but this layer will be removed if other models perform better without it.

Different models were tried out for image classification tasks. A logistic model based on the mean value of converted image arrays did not perform well, but CNNs with two and three layers increased performance. However, the ResNet-37 model, which includes convolutional layers, batch normalization, dropout, and global average pooling, was implemented to further improve performance and effectively address **vanishing gradients**. For predicting cancerous cells, the VGG model is a good choice due to its state-of-the-art performance, transfer learning capabilities, efficient use of computational resources, and ability to work well with small input sizes. The VGG model's architecture is suited for learning meaningful features from images, and its use of small **3x3 convolutional filters** allows for deeper networks with fewer parameters. After testing the VGG model with multiple blocks, the VGG model with 3 blocks was selected as

it was showing the best performance, and it can be fine-tuned with a small dataset, making it a good choice for this specialized task. (See Appendix A for the accuracy result).

The VGG 3 Blocks model was chosen for further tasks due to its superior performance compared to other models. **Hyperparameter tuning** was conducted using **keras tuner**, experimenting with different optimizers and epoch values. Overfitting was observed, but changing the compiler to SGD and implementing early stopping did not yield significant improvements. **Extra dropout layers** and **data augmentation** techniques were introduced during training to combat overfitting, which included random rotations, brightness adjustments, zooming, shifts, shearing, flips, and pixel filling. These augmentations aimed to diversify and expand the training data to potentially enhance the model's overall performance and generalization capabilities. (See Appendix B for the final model result).

Based on the hyperparameter tuning process, the best hyperparameters for the model were found to be a lambda value of None, a learning rate of 0.0001, and 25 epochs. Based on our **evaluation** the final model achieved a high accuracy of 90.8% with less overfitting and an F1-score of 0.51 in which the model is able to predict 58% of non cancerous cells and 44% of is cancerous cells correctly. This model is considered the best among the tuned models. However, there is room for improvement in the F1-score since the data is imbalanced. To address this, one can employ techniques such as resampling the data or using class weights to balance the dataset.

Task 2: Classifying Images according to cellTypeName

The **baseline model** consists of a flatten layer, a dense hidden layer with sigmoid activation, and a dense output layer.It provides a flexible structure for preprocessing input data, particularly image data, for multiclass classification tasks.

For predicting cellTypeName, we used the same VGG model as it was the best model for predicting isCancerous. We wanted to see if the VGG model also performs well on predicting cellTypeName. Additionally, we used three new models to test their performance on predicting cellTypeName. One of the models used for predicting cellTypeName was an ANN model. ANNs are well-suited for image classification tasks as they can extract meaningful features from images and learn complex non-linear relationships between the input features and output labels. To optimize the ANN model's performance, multiple hidden layer configurations were tested, and it was found that a 3 hidden layer ANN model provided the best accuracy for predicting cellTypeName from the given image data. In addition to the ANN model, we also tested the performance of three other models on predicting cellTypeName.

We chose to implement RCCNET based on its superior performance compared to VGG, AlexNet, and GoogLeNet as mentioned in the **referenced paper**[5]. RCCNET consists of four convolutional layers, two max pooling layers, and two flatten layers. However, despite our efforts, RCCNET did not perform as well as the VGG 2 Blocks model in our experiments. (See Appendix C for the accuracy result).

The VGG 2 Blocks model was chosen due to its superior performance compared to other models including other numbers of blocks of the same VGG model. **Hyperparameter tuning** was conducted using **keras tuner** same as task 1. Despite our efforts, applying **early stopping** changing the learning rate and lambda the model did not yield significant improvements**.Data augmentation** techniques were introduced to **reduce overfitting**. (See Appendix D for the final model result).

Based on the hyperparameter tuning process, the best model was the VGG 2 Blocks model with learning rate 0.0001.The model performed well in predicting cell type 0 with 44% accuracy and it gives overall accuracy of 72.7% and F1 score of 31%. This model is considered the best among the tuned models. However, there is room for improvement in the F1-score since the data is imbalanced as we have seen in EDA that epithelial is completely dominant in the dataset. To address this, one can employ techniques such as resampling the data or using class weights to balance the dataset.

Extra + Main

To include cellType attributes for the extra dataset, we utilized the CTFinal model. First, the CTFinal model was used to predict cellType and cellTypeName for the extra data. Then, we set the threshold of the **confidence level** given by the prediction at the 50th percentile, removing the lower half of the confidence level and adding only the confident data to the main dataset. This approach ensured that the CTFinal model, which was trained using the main dataset, was also able to learn from the extra dataset. Once the main and extra datasets were combined, we trained the CTFinal model again using the same parameters, resulting in the CTFinal_extra model. (See Appendix E for the final model result).

We evaluated our best model using the new CTFinal_extra data and observed overfitting. To address this issue and enhance model performance, we implemented data augmentation techniques. As a result, we achieved a maximum accuracy of 63.5% while the original CTFinal model achieved that of 51.3%, indicating an **improvement** by 12.12%p in the model's ability to generalize and combat overfitting.

Conclusion

Overall, the VGG model demonstrated superior performance in predicting both isCancerous and cellTypeName. Its effectiveness in training small pixelated images can be attributed to the utilization of small 3x3 convolutional filters. This key parameter played a crucial role in achieving outstanding results for these tasks. Therefore, the VGG model is deemed highly suitable for such image classification tasks. In our **independent evaluation**, we found that our model had a significantly lower F1 score compared to the models in the referenced papers. This discrepancy may be attributed to the dataset size and data imbalance observed during exploratory data analysis. To address the issue, we suggested upsampling the data as a potential solution, which was not covered in this paper.

References

1.)He, K. *et al.* (2015) *Deep residual learning for image recognition*, *arXiv.org*. Available at: https://arxiv.org/abs/1512.03385 (Accessed: 13 May 2023).

2.)Basha, S.H.S. *et al.* (2019) *Impact of fully connected layers on performance of convolutional neural networks for Image Classification*, *arXiv.org*. Available at: https://arxiv.org/abs/1902.02771 (Accessed: 13 May 2023).

3.)Simonyan, K. and Zisserman, A. (2015) *Very deep convolutional networks for large-scale image recognition*, *arXiv.org*. Available at: https://arxiv.org/abs/1409.1556 (Accessed: 16 May 2023).

*4.)Locality sensitive deep learning for detection and ... - IEEE xplor*e. Available at: https://ieeexplore.ieee.org/document/7399414 (Accessed: 16 May 2023).

5.)Shabbeer Basha, S H et al. “RCCNet: An Efficient Convolutional Neural Network for Histological Routine Colon Cancer Nuclei Classification.” 2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV). Ithaca: IEEE, 2018. 1222–1227. Web.

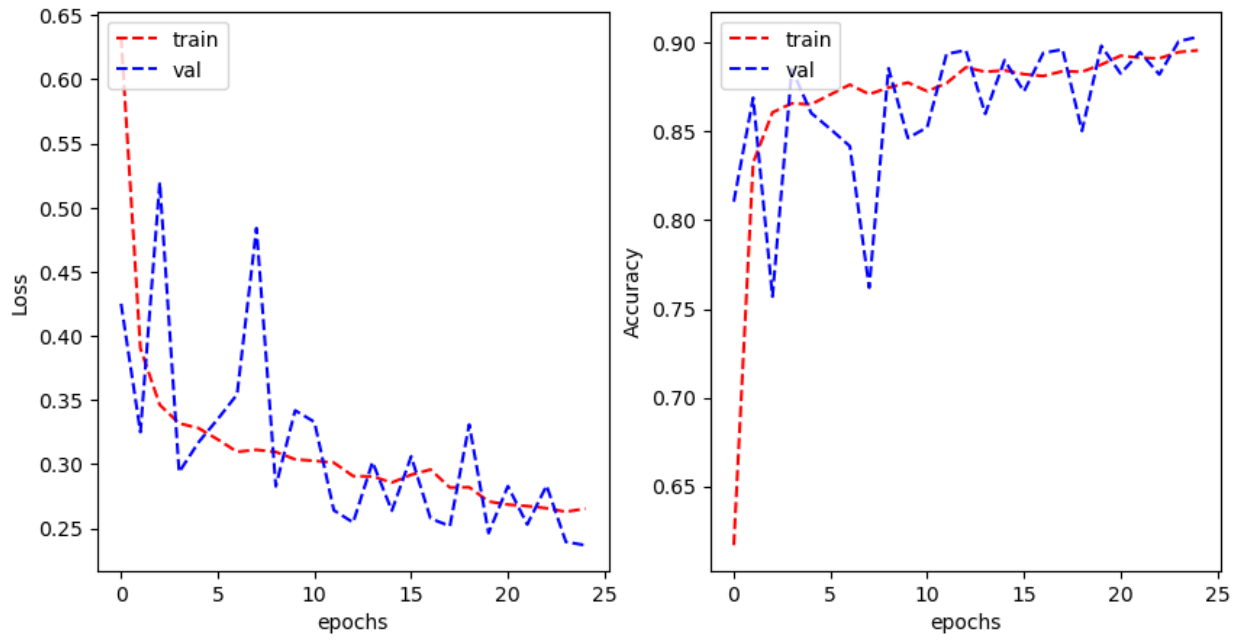
6.)Uppal, K. (2021) *Paper summary: ImageNet classification with Deep Convolutional Neural Networks*, *Medium*. Available at: https://karan3-zoh.medium.com/paper-summary-imagenet-classification-with-deep-convolutional-neural-networks-41ce6c65960 (Accessed: 16 May 2023).

Appendix

Appendix A

Model for Task 1	**Accuracy** **Test Accuracy**	**Overfitness** **|Train_acc - Test_acc|**	**F1 score**
VGG 3 Blocks	91.4%	0.2%p	53%
CNN 4 Layers	84.4%	0.19%p	49%
Base	86.9%	0.62%p	52%
ResNet	75.6%	0.02%p	48%
Logistic Regression	4.83%	1.67%p	-

Appendix B

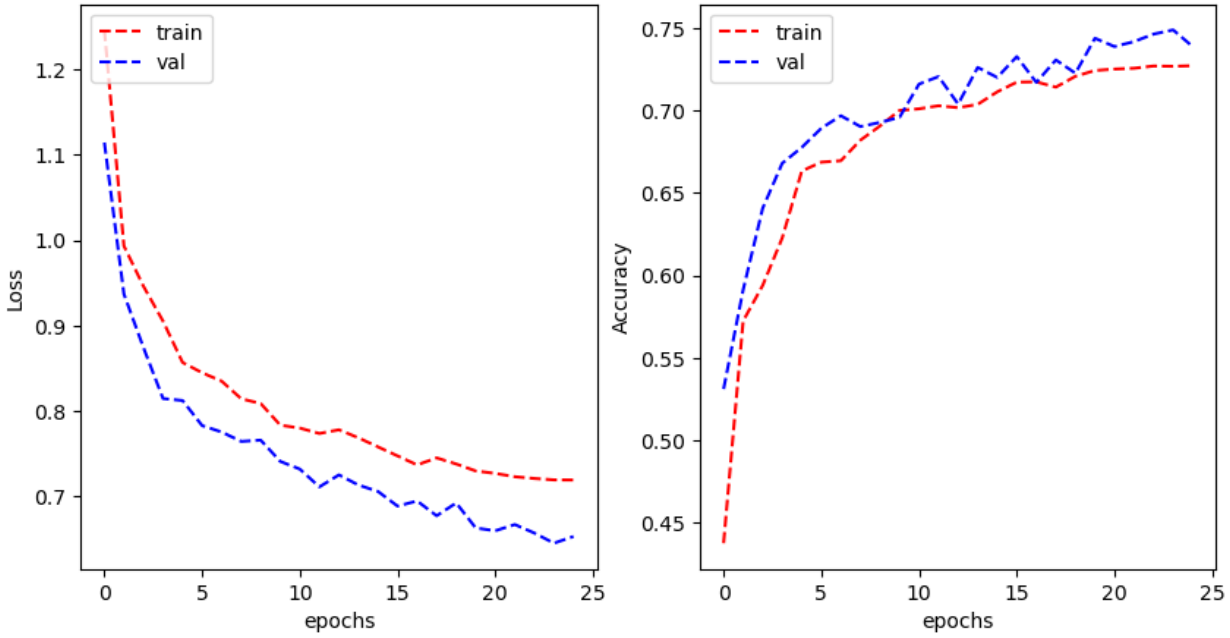


Appendix C

Model for Task2	**Accuracy** **Test Accuracy**	**Overfitness** **|Train_acc - Test_acc|**	**F1 score**
VGG 2 Blocks	75.5%	2.96%p	29%
ANN 3 Hidden Layer	69.9%	1.55%p	31%
Base	66.9%	2.68%p	30%

AlexNet	61.1%	1.19%p	32%
RCCNet	57.5%	18.71%p	26%

Appendix D



Appendix E