MLP Coursework 4

Group G102 (s1722201 & s1640380 & s1778742)

Abstract

This coursework presents a second approach to the tumor tissue image classification task. We implemented an Inception-V3 deep neural network that detects and classifies breast tumor cells over tissue slide images. Due to the difficulty of the problem, we further explored and compared several image pre-processing methods over different image magnifications, hypothesizing that simple pre-processing may be beneficial in this setting. These experiments allowed us to measure the impact of normalization, standardization, augmentation, segmentation and dataset balancing of medical tissue slides on our model. Despite the lack domain knowledge about the pathology of this disease, our neural architecture achieved 93.0% ± 0.5% accuracy for 40X magnification image slides. We found that unlike other methods which rely heavily on complex feature engineering, convolutional neural networks with simple image pre-processing techniques can achieve good results, comparable to results obtained using significantly more computationally intensive convolutional neural network models.

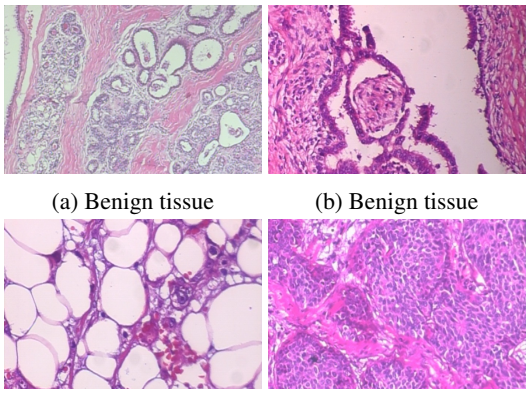
1. Introduction

In the early 90's, the field of computer vision moved from using edge detectors, lines, and filters to supervised learning techniques. This shift allowed the subfield of medical image analysis to develop trainable systems that facilitate the processing and analysis of medical images. The previous approach is characterized by training a system with custom-made patterns. Despite the fact that supervised methods are still widely used today, recent advances and results in deep neural networks research have led the medical image analysis community to consider neural networks.

Lo et al. (1995) was the first to apply neural networks to medical images. However, it was not until the late 1980s that LeCun et al. (1998) introduced LeNet to efficiently recognize character images. In 2012 Krizhevsky et al. (2012) implemented AlexNet, an increasingly complex neural architecture that won the ImageNet challenge. This background motivated us to explore deeper and more robust architectures on a set of medical image data.

1.1. Research Questions

In the previous coursework we presented a first approximation to the breast cancer image classification dataset (BreakHis)[1] compiled by Spanhol et al. (2016a). Briefly, as covered in Coursework 3, the data is in the form of images from slides of breast tissue to be classified into either Malignant or Benign samples. The 7915 images in the dataset are of 40x, 100x, 200x or 400x magnification and come from 81 patients. There are approximately 31 percent benign samples and 69 percent malignant samples.



(a) Benign tissue (b) Benign tissue

(c) Malignant tissue (d) Malignant tissue

Figure 1. Example images of both Benign and Malignant tissue, image taken from (Spanhol et al., 2016a)

In coursework 3 we implemented and compared the performance of three baseline models. More specifically, we used classical machine learning methods like support vector machines (SVM), and two deep neural network architectures, LeNet (LeCun et al., 1998) and GoogLeNet (Szegedy et al., 2015). We discovered that the best-performing baseline was GoogLeNet, which achieved 83% validation accuracy on average. We found that classifying tissue images for detecting benign or malignant presence of cells is a challenging task.

In this coursework we present a second approximation for the breast cancer image classification task. Based on the results presented in Szegedy et al. (2016) and coursework three, we proposed an Inception-V3 classification architecture for classifying malignant or benign tissue slides hypothesizing that the deeper network architecture would yield better results in this case. Following this, we investigated the effects of image pre-processing for this task,

[1] https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/

hypothesizing that better performance could be obtained by altering the images either in accordance to how human researchers may view these images, such as patch by patch at a time, or by changing other heuristics such as color. Due to time constraints, our primary focus was on pre-processing. Model selection was validated on the previous coursework, hyper-parameters like learning rate and momentum were initialized with 0.01, and 0.9 respectively. Hyper-parameter tuning could be easily done with more time and would not be as interesting as exploring the effects of simple pre-processing techniques. It has also been extensively done for other architectures in Courseworks 1 and 2.

The research questions for this coursework can be summarized as:

- What is the performance of a state-of-the-art neural network like Inception for this task in comparison to the best performing baseline from Coursework 3?
- What is the effect of normalization and standardization over our dataset?
- Can our model generalize well for different image magnification levels?
- What is the effect of data augmentation in our classification pipeline?
- Does applying segmentation techniques increase the performance of our pipeline?
- Does balancing the classes in the dataset when training the model improve the performance of the model?
- What is the overall effect of the data pre-processing techniques outlined above on the performance of this model and would there appear to be merit in further exploring data pre-processing?

Therefore the objectives are to investigate color standardization, color normalization, dataset augmentation, image segmentation and dataset balancing. The methodology itself is described in more detail in Section 3.

1.2. Dataset Preprocessing

We approached this problem as a supervised binary classification task. For the initial part of the experimental setup, there was no pre-processing involved, apart from splitting the images into Training, Validation, and Test Sets. The below table shows the final configuration of our data, after splitting it.

Type	Test set	Valid	Train
Benign	325	644	1514
Malignant	599	1093	3740
Total		7915	

Table 1. Data configuration.

We created a script that loads the images as a scaled numpy matrix, and given the directory or location (i.e., benign or malignant) of the sample, it automatically labels each image sample or row vector with 0 or 1. Additionally, the dataset was split by patient ID, which ensures that the samples are mutually exclusive. In other words, no samples from one patient are repeated across the training, testing, and validation set. This is consistent with the literature. The convolutional neural networks were run over the matrix of pixel values. This setting allowed us to explore the behavior of the models with raw data although the pixel values were divided by 225 to move values to the range 0 to 1 and avoid issues with numerical overflow which could otherwise arise, for example, in the non-linear activation layer of a neural network. Also, based on the previous coursework, we re-sized the images to 224x224.

2. Related Work and the Breast Tissue Classification Task

Previous work on the classification task for this dataset falls into two main categories; non-CNN approaches and CNN-based approaches. Both use the same overall method of splitting the data into training and test sets without having individual patients belonging to both sets.

The dataset was introduced by Spanhol et al. (2016a) who investigated combinations of 6 state-of-the-art texture representation techniques including local binary patterns (Guo et al., 2010) and parameter-free threshold adjacency statistics together (PFTAS) with four types of classifiers; one nearest neighbor, linear quadratic analysis, random forests and Support Vector Machines. The models were trained separately for each magnification level. The best result obtained was by the one-nearest neighbor or Support Vector machine classifiers using the PFTAS descriptor. The error bars are relatively high, and the results vary by magnification level. They identified particular issues with one type of benign tumor, fibroadenoma, which was similar in structure to a malignant tumor, leading to relatively high false positive rates. Another negative aspect of these traditional methods is that they heavily rely on complex feature engineering techniques.

Building on this work, Spanhol et al. (2016b) proposed the use of CNNs due to their state-of-the-art performance on many image classification tasks. Obtaining approximately 90% classification accuracy on the test set for 40x magnification, 88% for 100x, 85% for 200x and 86% for 400x by training a separate model for each magnification level. They used random patch extraction and the AlexNet architecture (Krizhevsky et al., 2012), improving on the previous work by 4-6% by combining the results of 4 CNNs using different segmentation methods with the maximum rule. A disadvantage of this work is the computational cost of the AlexNet architecture, as discussed again in Section 3.1. The research of Spanhol et al. (2016b) also mentions that histopathologic image classification, in general, is a challenging task, mainly because images present a high variance of rich and complex geometrical structures.

Spanhol et al. (2017) then investigated a method in between these two previous approaches which reused a previously trained CNN, CaffeNet, and extracted the output of the final layers of this model as the feature vectors then fed to a logistic regression classifier. The results were comparable to those of Spanhol et al. (2016b), although did not outperform them.

Finally, Bayramoglu et al. (2016) attempted to build a CNN classifier independent of the magnification level as they state that is would not always be practical to need images of a specific magnification. They point out the high level of variability in the samples may be due to many factors such as different lab protocols, different staining protocols and different levels of skill of the person preparing the tissue sample. They also performed data augmentation via some basic rotations of 90, 180 and 270 degrees, flipping images and cropping them to 460x460. While their results do not outperform those of Spanhol et al. (2016b), they suggest the investigation of stain normalization, deeper architectures and more training data going forward.

3. Methodology

Contrary to the previous coursework, all models and experiments were implemented in a Python environment. We used Keras (Chollet et al., 2015), a higher level interface of TensorFlow (Abadi et al., 2016). We decided this time to focus just on a Keras implementation because we found it to have have faster prototyping and the purpose of this coursework was to experiment with pre-processing and not tuning the model in detail.

Additionally, we setup our own Google Cloud cluster with 52 GB RAM, 4CPU core, and 1 Nvidia K80 GPU board. This allowed us to run several experiments in a short period. This was particularly crucial as all tests were run 3x or 5x each to obtain an estimate of the variability of the result and the dataset was large.

3.1. Inception-V3 Architecture

For this coursework, we used a convolutional neural network architecture named InceptionNet, an advanced version of GoogLeNet. In a traditional convolutional layer, a filter of a given size, say 5x5 is 'scanned' over the input image. This filter has the same weight matrix at all points in the image and therefore is essentially searching for an individual learned feature, such as an edge, across patches of the entire input image. There can be multiple input channels, and multiple filters applied. The hidden layer obtained from the input for a convolutional layer is defined typically for a given feature as:

$$\mathrm{h} = \mathrm{sigmoid}(\mathbf{w} \bigotimes \mathbf{x} + b)$$

where h is the hidden layer, $\mathbf{w}$ is the weight matrix, $\mathbf{x}$ is the input, b is the bias term, and $\bigotimes$ indicates cross-correlation.

The resulting matrix of values for that feature is then pooled into a smaller representation using a pooling mechanism such as MaxPooling. Max-pooling simply takes the maximum value of units in a region, for example taking the maximum value of each 2x2 region in the hidden layer to form a smaller hidden layer. These layers are combined with fully connected layers using non-linear activation functions such as the Rectified Linear Unit (ReLU) (Nair & Hinton, 2010) and an output Softmax Layer transforming the output to a probability distribution. These layers then form a network which can then be trained using, for example, stochastic gradient descent. The following shows an example architecture of a 'classic' CNN:

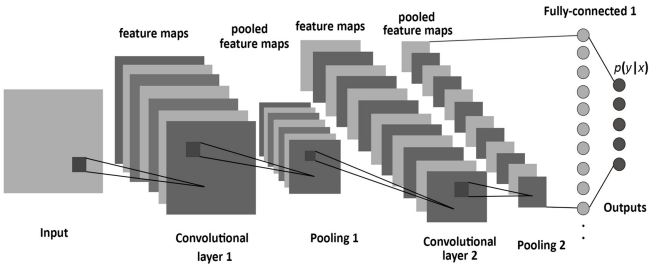


Figure 2. Traditional CNN framework from (Albelwi & Mahmood, 2017).

InceptionNet aimes to reduce the high computational overhead of convolutional layers used in the top performing convolutional architectures by breaking relatively large convolutions into a network of smaller convolutions, using their inception module framework:

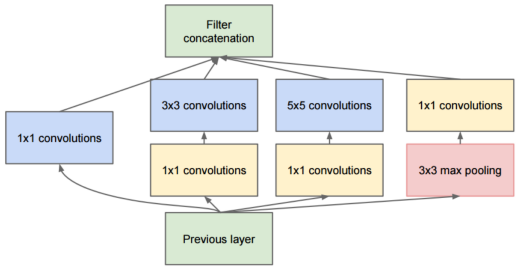


Figure 3. Inception module, image taken from (Szegedy et al., 2015).

These modules essentially factorize larger convolutions into several smaller ones in a computationally efficient manner with the goal of capturing the same degree of expressivity as a larger convolution. The 1x1 and 3x3 convolutions and max-pooling are performed first to reduce the cost of the expensive 5x5 and 3x3 convolutions which follow. This network of smaller convolutions has overall fewer parameters.

These modules are then combined into a 22-layer architecture as shown in Figure 3.

The use of these modules reduces the number of parameters from 60 million in the popular CNN architec-

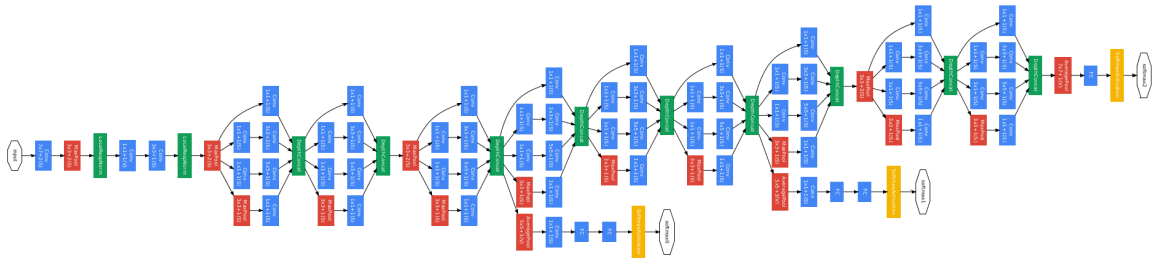


Figure 4. GoogLeNet framework from (Lazebnik, 2017). The blue nodes are convolutional layers, red are pooling layers, yellow are softmax layers, and green are other such as concatenation layers

ture AlexNet (Krizhevsky et al., 2012) to 5 million in GoogLeNet (Szegedy et al., 2015) while maintaining performance. In this part of the report, the successor of GoogLeNet, the Inception-V3 (Szegedy et al., 2016) architecture is used.

The Inception-V3 architecture aims to scale up the representative power of GoogleNet without losing the advantage of the reduced computational complexity in comparison to other deep convolutional architectures such as AlexNet. It essentially adds the following features to GoogLeNet to allow for increased model depth and representative ability without an unfeasible increase in computational cost:

1. Factorizing 7x7 convolutions, into structures similar to the inception module using convolutions of size 3x3, 1x1, 7x1 and 1x7
2. Label Smoothing is used as a method of regularizing the classification layer by essentially adding uncertainty by estimating the marginal effect of label dropout on the resulting classification
3. The addition of Batch Normalization (Ioffe & Szegedy, 2015) to the auxiliary classifiers used in GoogLeNet (the classifiers which are attached at the bottom of the framework shown in Figure 3) to further improve their performance as regularizers

The architecture also implements a transfer learning approach which initializes the network with weights pretrained using the ImageNet dataset (Deng et al., 2009). This allows increased training speed. The culmination of these improvements is that Inception V3 outperforms GoogLeNet on the ImageNet task. It also exceeds GoogLeNet on this task as discussed in Section 4 and thus is used for the remainder of this report.

3.2. Normalization

The images in the dataset show signs of having different levels of staining (Bayramoglu et al., 2016), potentially due to different amounts of dye or other protocols being used in different labs collecting the tissue. For this reason, we decided to normalize our dataset for color variation. Thus, we tried three different approaches:

1. **Gray Scaling**: The images were changed to greyscale
2. **Contrast Normalization**: Contrast Limited Adaptive Histogram Equalization (CLAHE) (Zuiderveld, 1994) enhances the local contrast of an image. This was implemented as it has been shown to improve the performance for other tasks involving images of breast tissue (Pisano et al., 1998)
3. **Channel Standardization**: The pixel values over each channel were standardized to 0 mean and unit variance to have all inputs in a similar range of values

3.3. Augmentation

According to Wang & Perez (2017), data augmentation can help to increase the performance of an estimator, the authors compared the effect of several data augmentation techniques like cropping, rotating, and flipping input images. Based on the previous research paper (Spanhol et al., 2016b) we augmented the data by flipping left to right, flipping top to bottom, and 90°/80° rotation as well as by performing light variation by adding 10% color distortion.

3.4. Segmentation

Segmentation was performed by splitting the images into 2x2, 3x3 or 10x10 patches. Havaei et al. (2017) present a deep learning method that performs brain tumor segmentation. One of the positive aspects of this approach is that the system learns to identify relevant regions of interest that are specific to detect brain tumors. Given the fact that our images presented significant irrelevant sections (e.g., fat, healthy skin, and pigmentation areas), we decided to run image segmentation to eliminate irrelevant image portions over our image slides.

In the related work for this dataset, the simple approach of extracting random patches of images was used. We also decided to implement a simple method to assess if removing patches helped our model, intending to increase the complexity of the approach if initial results were promising. The segmentation approach was motivated by the fact that large parts of images contained regular tissue instead of the cells which contained the relevant information for cancer diagnosis for a human observer. These less rel-

evant patches were thought to be characterized by being of lighter color, without dark patches of cells. Therefore, thresholding was first applied to the images, with all values above the threshold set to 1, therefore identifying the light pixels. The pixels of each patch were then summed, and the patch with the largest value kept for classification. By manual analysis this appeared to work as expected, extracting a patch containing darker parts which resembled areas of cells. An example is included in section 4.6.

3.5. Balancing

In the previous coursework, it was discovered that our classes were unbalanced with around 70% malignant samples in the training dataset. Although this is to be expected as biopsies are only taken of tumors which may be malignant, we decided to investigate the impact of balancing our dataset as it is unclear the exact effect this imbalance would have on model performance. More specifically, since the model is heavily biased towards malignant class, making up 70% of the training data and as we do not know the posterior probability of having malignant presence of cells, we decided to balance our dataset by augmenting flipped left to right images examples, this resulted in a 50/50 proportion of benign and malignant examples.

4. Experiments and Results

In this section, we present the experiments and results for our inception based image classifier. In the previous coursework, we propose three different baselines. However, on average our best model was GoogleNet because it showed 83% average validation accuracy. Inception-V3 is used here as it is the latest version of GoogLeNet which had been shown to outperform GoogLeNet on the ImageNet tasks as described in Section 2. This model is run first without any pre-processing to provide a direct comparison to GoogLeNet.

4.1. Experiment List

The following experiments were run, all using the Inception-V3 architecture:

1. Inception-V3 (IV3)
2. IV3 + Greyscale + Channel Standardization
3. IV3 + Greyscale only
4. IV3 + Split Magnification
5. IV3 + Split magnification, greyscale and standardization
6. IV3 + Split Magnification and greyscale
7. IV3 + Split Magnification and CLAHE contrast normalization
8. IV3 + Split Magnification and augmentation except for color distortion
9. IV3 + Split Magnification and all augmentation
10. IV3 + Split Magnification with 2x2 Patch Segmentation
11. IV3 + Split Magnification with 3x3 Patch Segmentation
12. IV3 + Split Magnification with 10x10 Patch Segmentation
13. IV3 + Split Magnification with Balanced Malignant and Benign classes

4.2. Results

The following results were obtained for experiments 1-3, each run 5 times. The below table shows the model's maximum validation accuracy:

Experiment	Max. Validation Accuracy +/- Two Standard Deviations
1	85.05% +/- 1.80%
2	81.19% +/-0.88%
3	80.11% +/- 1.02%

Table 2. Results table per image magnification.

It is notable that the Inception Architecture here outperformed the GoogLeNet architecture used in coursework three which attained an average validation accuracy of 83.3% ± 0.5% over five runs. Therefore the Inception framework is used going forward; potentially the deeper architecture allowed slightly more complex relationships to be captured by the model.

Another notable aspect is the following confusion matrix:

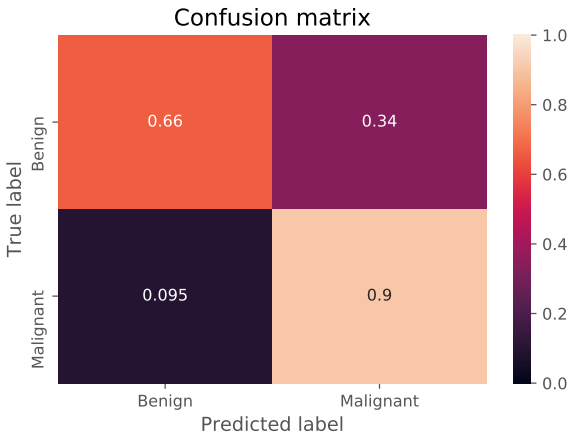


Figure 5. Confusion matrix for experiment 1

As we can see in the above image, the false positive rate is very high. The following results were run three times for each experiment, and the results averaged. Interestingly, we found that the model had better performance with images at 40X magnification than any other by a significant margin.

Experiment	40X	100X	200X	400X
4	90.30% +- 0.15%	85.32% +-0.77%	82.73% +-2.39%	81.89% +-1.08%
5	90.21% +-0.53%	82.77 +-1.85%	74.38% +- 1.22%	68.82% +- 0.87%
6	90.30% +-0.73%	77.30% +-8.45%	76.09% +-0.87%	70.39% +- 3.06%
7	89.11% +-1.77%	84.26% +-1.49%	82.30% +-0.49%	82.05% +- 2.72%
8	89.79% +-0.77%	86.24% +-1.07%	85.37% +-0.69%	81.64% +-1.51%
9	92.57% +-0.29%	82.48% +-1.97%	80.73% +-1.19%	81.97% +- 1.03%
10	84.14% +-1.52%	82.48% +-1.60%	82.44% +-1.13%	79.98% +-1.88%
11	80.93% +-1.05%	80.28% +-0.25%	81.23% +-2.71%	80.23% +-1.27%
12	80.25% +-0.67%	78.09% +-0.98%	78.66% +-0.65%	76.26 +-1.00%
13	93.00% +-0.53%	83.83% +-0.64%	82.51% +-1.62%	78.49% +-3.33%

Table 3. Results table per image magnification.

The results are discussed in depth in the following sections.

4.3. Split Magnification

For this experiment we trained a model per each image magnification, motivated by the fact that the validation accuracy for the Inception model was converging very noisily, as shown in the following plot:

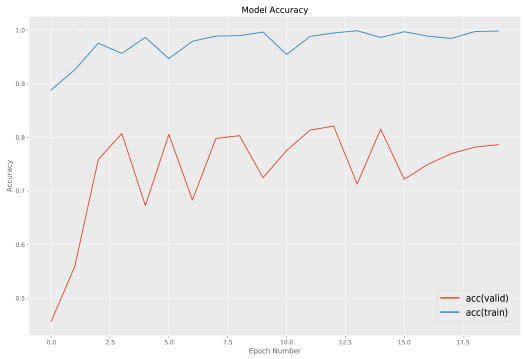


Figure 6. Validation Accuracy for Overall Inception Model

We conjectured this might be due to the different magnification images providing conflicting information and features, we, therefore, split the images by magnification level and trained a separate model on each. This improved the convergence of the validation accuracy as shown in the following plot:

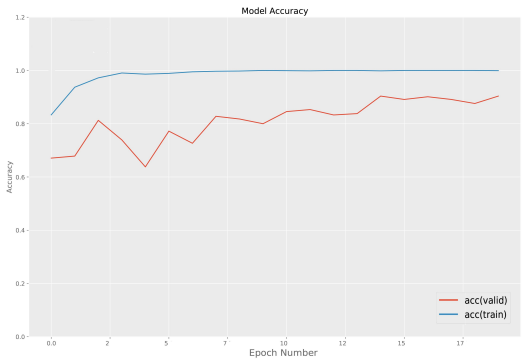


Figure 7. Validation Accuracy for 40x Magnification

The differences in performance indicate that each magnification level is its own classification task with its difficulty level and potentially its own features to be learned. This intuitively makes sense as looking at the 40x image, a lot more can be said about the overall structure of the tissue but looking at 200x magnification more can be said about cell shape. These features may have different levels of relevance to the classification task at hand.

Additionally, the confusion matrices show an improvement in the level of false positives when compared to the overall model:

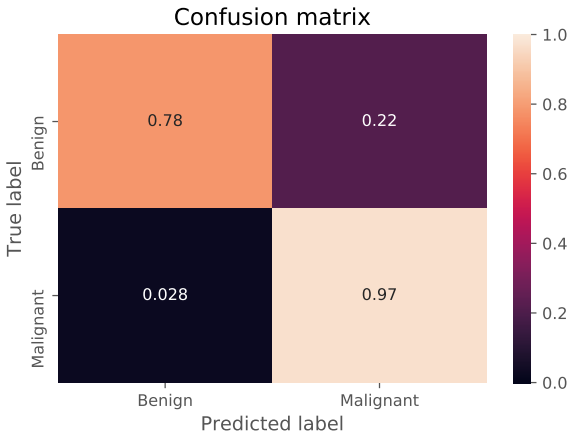


Figure 8. Confusion matrix for experiment 4 at the 40x magnification level

This was true for all magnification levels except for 400x which had a worst case of 37% false positive rate for benign samples.

4.4. Normalization and Standardization

Neither normalization nor standardization improved the validation accuracy. This may be due to our assumption that different levels of dye are due to differences in lab practices. For example, it could, in fact, be due to some unseen use of the different levels of color for example if the surrounding cells of cancerous tissue absorb more dye and the heightened color can thus be used as a useful feature. CLAHE normalization also did not improve the validation accuracy. However, this experiment in particular would ideally have been subject to further hyper-parameter tuning, time permitting. This was not pursued due to the time restrictions as the reported results were not particularly promising.

4.5. Augmentation

Augmentation without color distortion attained the best overall validation accuracy for 100x and 200x magnification levels and only slightly decreased validation accuracy for 40x and 400x. It is the best model on average over all magnifications. It suggests that availability of more training data would help the performance of these models as in this case rotated images are very similar to true

images. Adding images with color distortion negatively effected performance for the 100x and 200x magnification levels, but had little effect for the 40x and 400x magnification levels when the size of the standard deviation is taken into account. This reinforces the conclusion from the color normalization experiments that the color is potentially an important feature.

4.6. Segmentation

The results of the segmentation experiment did not improve the overall accuracies attained. However, we feel that despite this the results were quite promising considering the naive method used. Although no magnification improved, there was also not as much performance loss as might have been expected. This suggests that further experiments using more complex segmentation approaches could yield promising results.

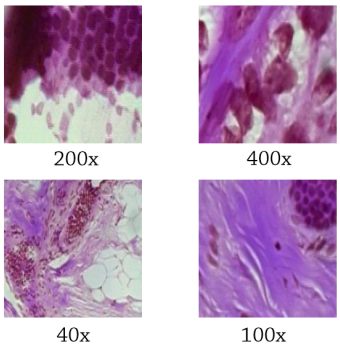


Figure 9. The segmentation approach for several magnifications.

The above image shows the output for our segmentation approach. Interestingly, we can see that malignant cell clusters are detected efficiently. Other, segmentation techniques like structural or stochastic may require more time to tune.

4.7. Balancing

Balancing classes achieved the highest validation accuracy attained for the 40x magnification level. It also improved the false positive rate for this magnification level.

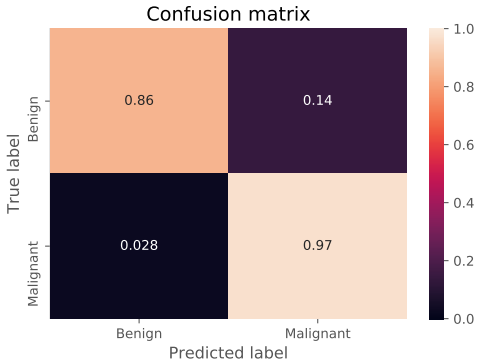


Figure 10. Confusion matrix for experiment 14 at the 40x magnification level

However, class balancing did not show similar improvements for the other magnification levels. The combination of this result with those of the other differing experiments again suggests that each magnification level is an entirely different problem as the class split and therefore the class imbalance is similar for each magnification level.

In fact, class balancing very negatively affected the false positive rate of the 400x magnification level:

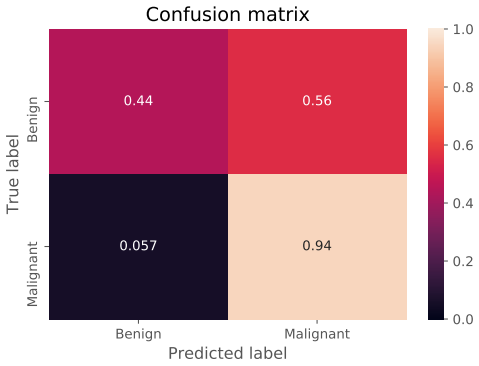


Figure 11. Confusion matrix for experiment 14 at the 400x magnification level

4.8. Results Comparison to Related Work

The best related work results were obtained by Spanhol et al. (2016b) who had a similar result where different CNNs, in their case CNNs with different segmentation techniques, performed best on different magnification levels. Their evaluation was performed on their validation set as they did not have a separate test set, therefore the below table compares our validation set performance for the best model for each magnification level to theirs, however the difference in the size of the split of 20% validation set in this report versus 30% validation set in their report makes the comparison challenging:

Model	40x	100x	200x	400x
Best Model from Spanhol et al.	89.6% +- 6.5%	85.0% +- 4.8%	84.2%+-1.7%	81.6% +- 3.7%
Best Model from this work	93.0% +-0.5%	86.2%+-1.1%	85.4+-0.7	82.1+-2.7%

Figure 12. Validation Accuracies Comparison

Our best CNN for each magnification level has a higher average validation accuracy, but also a significantly lower standard deviation making it difficult to make a direct comparison.

4.9. Test Set Results

Finally, the best model for each magnification level was chosen based on the results above. These models were then evaluated using the held out test set to obtain an estimate of their generalization ability. The accuracy for 40x was 87.07%, 100x was 90.00%, 200x was 94.57% and 400x

was 64.84%. These results are relatively unexpected, particularly the inferior performance for the 400x magnification level. We suspect that the test set and potentially also the validation set sizes are too small given the complexity of the training data.

The receiver operating characteristic curves (ROC) were then plotted, illustrating the ability of our inception V3 network to classify images slides correctly. This helped to contrast the actual positive rate against false positive rate. Again, the 400X image magnification model was the worst performing model in this respect.

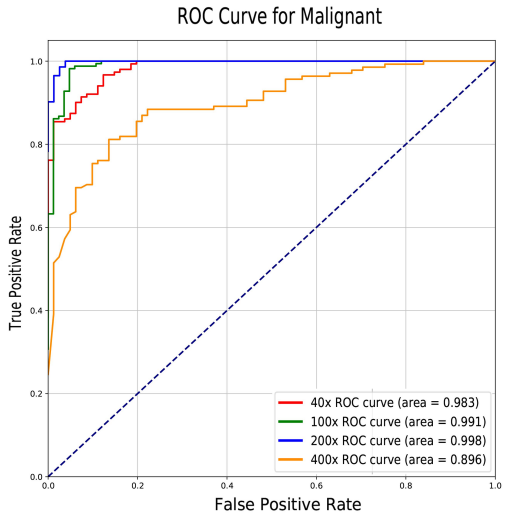


Figure 13. ROC curves over the test sets.

Finally, the below image corresponds to the confusion matrices of the model over the test set using the model with best validation accuracy.

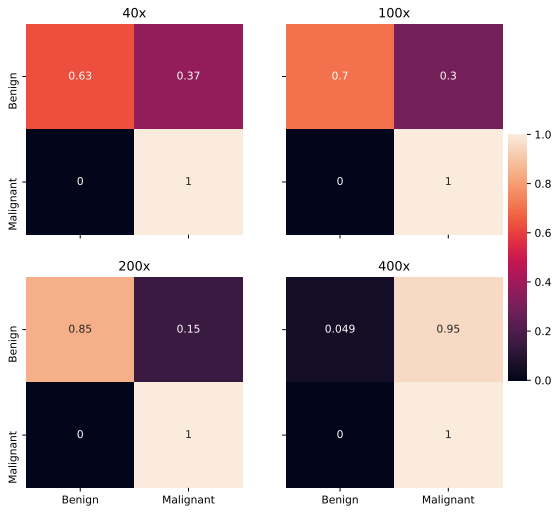


Figure 14. Model evaluation over test sets evaluations.

The confusion matrices for the 40x, 100x and 200x magnification levels are in line with what could be expected given the confusion matrices for these magnification levels on the validation set. The 400x model is predicting most samples as malignant. This was not the case for the validation data and is a puzzling result.

5. Conclusions

This coursework presented a second approximation to the breast-cancer-image-classification dataset (BreakHis); we found that separating per magnification improved classification performances. Therefore, each magnification level presents different problems as some magnification benefits from particular pre-processing technique but not the other. Another interesting insight was that balancing our data played an important role in increasing the performance of our system for one magnification level, while dataset augmentation benefited other magnification levels. To summarize, the main outcomes of this coursework were the following:

- We experimented with a state-of-the-art deep neural architecture.
- We successfully set up a Google Cloud environment which let us run multiple complex models and several experiments in a short amount of time.
- We successfully implemented an Inception V-3 pipeline.
- We experimented with the impact of normalization our dataset, and we investigated the behavior a complex model over different image sizes.
- We researched and tested image augmentation techniques.
- We developed an image segmentation pipeline to detect malignant tissue clusters.

As a future work, we would like to explore and compare the performance of deeper models like densely connected convolutional networks, which are networks that connect each layer in a feed-forward approach. Huang et al. (2017) state that these models have several advantages, for example, these models can compensate the vanishing-gradient problem, strengthen feature propagation, encourage feature reuse, and substantially reduce the number of parameters. Another interesting architecture that we would like to try is VGG nets (Simonyan & Zisserman, 2014); these networks are well known for its depth and their relatively small 3x3 convolutional filters. VGG nets achieved good performance at ImageNet's classification and localization tasks.

Another area that would ideally be further explored is hyper-parameter optimization. For example, we would like to investigate what is the optimum size of the window as well as further investigate some parameters of Inception, such as the learning rate. Improving performance in this way may be possible. Finally, we can also explore advanced segmentation techniques by using pre-trainned regional proposing CNN to detect individual cells, such as the research work presented in (Johannesu, 2017).

References

Abadi, Martín, Barham, Paul, Chen, Jianmin, Chen, Zhifeng, Davis, Andy, Dean, Jeffrey, Devin, Matthieu, Ghemawat, Sanjay, Irving, Geoffrey, Isard, Michael, et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, volume 16, pp. 265–283, 2016.

Albelwi, Saleh and Mahmood, Ausif. A framework for designing the architectures of deep convolutional neural networks. *Entropy*, 19(6), 2017. URL http://www.mdpi.com/1099-4300/19/6/242.

Bayramoglu, N., Kannala, J., and HeikkilÃd', J. Deep learning for magnification independent breast cancer histopathology image classification. *23rd International Conference on Pattern Recognition (ICPR)*, 2016.

Chollet, François et al. Keras, 2015.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Guo, Zhenhua, Zhang, Lei, and Zhang, David. A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing*, 19(6):1657–1663, 2010.

Havaei, Mohammad, Davy, Axel, Warde-Farley, David, Biard, Antoine, Courville, Aaron, Bengio, Yoshua, Pal, Chris, Jodoin, Pierre-Marc, and Larochelle, Hugo. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.

Huang, Gao, Liu, Zhuang, Weinberger, Kilian Q, and van der Maaten, Laurens. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, number 2 in 1, pp. 3, 2017.

Ioffe, Sergey and Szegedy, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Bach, Francis and Blei, David (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL http://proceedings.mlr.press/v37/ioffe15.html.

Johannesu. Cnn cell detection. https://github.com/johannesu/cnn-cells, 2017. Accessed: 2018-03-26.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc., 2012.

Lazebnik, Svetlana, 2017. URL http://slazebni.cs.illinois.edu/spring17/lec01_cnn_architectures.pdf.

LeCun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lo, S-CB, Lou, S-LA, Lin, Jyh-Shyan, Freedman, Matthew T, Chien, Minze V, and Mun, Seong Ki. Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Transactions on Medical Imaging*, 14(4):711–718, 1995.

Nair, Vinod and Hinton, Geoffrey E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 807–814, USA, 2010. Omnipress. ISBN 978-1-60558-907-7. URL http://dl.acm.org/citation.cfm?id=3104322.3104425.

Pisano, Etta D., Zong, Shuquan, Hemminger, Bradley M., DeLuca, Maria, Johnston, R. Eugene, Muller, Keith, Braeuning, M. Patricia, and Pizer, Stephen M. Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *Journal of Digital Imaging*, 11(4):193–200, 1998. ISSN 0897-1889. doi: 10.1007/BF03178082.

Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. corr abs/1409.1556 (2014). *arxiv. org/abs/1409.1556*, 2014.

Spanhol, Fabio A, Oliveira, Luiz S, Petitjean, Caroline, and Heutte, Laurent. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016a.

Spanhol, Fabio A, Oliveira, Luiz S, Cavalin, Paulo R, Petitjean, Caroline, and Heutte, Laurent. Deep features for breast cancer histopathological image classification. In *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*, pp. 1868–1873. IEEE, 2017.

Spanhol, Fabio Alexandre, Oliveira, Luiz S, Petitjean, Caroline, and Heutte, Laurent. Breast cancer histopathological image classification using convolutional neural networks. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pp. 2560–2567. IEEE, 2016b.

Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent, Rabinovich, Andrew, et al. Going deeper with convolutions. *ICML*, 2015.

Szegedy, Christian, Vanhoucke, Vincent, Ioffe, Sergey, Shlens, Jon, and Wojna, Zbigniew. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Wang, Jason and Perez, Luis. The effectiveness of data augmentation in image classification using deep learning. Technical report, Technical report, 2017.

Zuiderveld, Karel. Graphics gems iv. In Heckbert, Paul S. (ed.), *Graphics Gems IV*, chapter Contrast Limited Adaptive Histogram Equalization, pp. 474–485. Academic Press Professional, Inc., San Diego, CA, USA, 1994. ISBN 0-12-336155-9. URL http://dl.acm.org/citation.cfm?id=180895.180940.