Investigations on Data-Centric Machine Learning for Medical Images and Signals

Introduction

This project attempts to explore the correlations between different data augmentation techniques on cancer classification and the combinations of different augmentation approaches and their impact on model accuracy. The thought behind these experiments is building a *robust model* through 1) training set augmentations that I think would help the model generalize better, 2) combining them in a helpful fashion that makes our model resistant to domain changes, then 3) trying to emulate some of the ideas in class related to dropout through augmentations instead of through model structure.

Methods

Experiment 1: Individual Data Augmentations

For the first section of the experiments, 10 data augmentation techniques are selected: 2 being orientation techniques (*horizontal flipping* and *rotation*), 4 being color-change techniques (*hue*, *saturation*, *contrast*, and *gamma changes*), 2 being noise-related techniques (*sharpness* and *blurring*), and 1 being a structure transformation technique (*shearing*) **[1]**. An individual training rotation of each of these techniques is preformed: for a subsampled dataset, a model is trained on the augmented version of the data for each of the 10 augmentations selected. Arguments/parameters per augmentation are given in **Figure 2** and **Figure 3**. Going forward, the *TensorFlow.Image* library **[2]**, the *Tensorflow_Addons* library **[3]**, and the

Tensorflow.Keras.Preprocessing.Image library [4] are used to produce augmentations. And the *Matplotlib* library [5] is used to produce training/accuracy graphs.

A subsampled dataset is gathered by randomly selecting 450 images from each class from the extracted 90x90 pixel cells of the ROI images provided [6], totaling 1350 images in total. For each augmentation selected, an individual model is trained on the same subset of 450 images referenced above. The training and validation accuracy history for each model over 30 epochs is used as a reference for an augmentation's performance. The AUROC score is used as the accuracy metric [7]. Maximum validation accuracy history per model is averaged over two runs to account for variability due to a small training set size. The same experiment is performed with a larger subset of training points of 900 images per class, totaling 2700 images in total. After the results from this first experiment are gathered (see **Figure 2** and **Figure 3**), the augmentations are categorized on how well they performed in both the 1350 image and 2700 image runs. Augmentations are categorized into the *'safe'* category if they perform better than non-augmentation in maximum validation accuracy in both runs, the *'watchlist'* category if they perform better than maximum validation accuracy in at least one of the two runs, and the *'dangerous'* category otherwise.

Experiment 2: Creating Augmentation Pipelines

Augmentations categorized into the safe category are combined into a pipeline called the '*safe approach*'. When a model is trained with this augmentation function, all augmentations in the safe category are performed on all training images. This safe approach is compared with a non-augmented approach and other '*watchlist approaches*' which are the fusion of the safe approach with some augmentation from the watchlist category added in. Too prevent

augmentations from drowning out real examples, the augmentation pipeline is weighted by which augmentations perform best. Safe operations are always preformed, but watchlist operations have a 25 percent chance *each* of not being performed. This experiment is the same logistically to **Experiment 1** in regard to how data is selected, the number of epochs, and what metrics are collected. The approach that performed the best in maximum validation accuracy times the number of augmentations (see **discussion** and **Figure 5**) was selected as the best approach moving forward.

Experiment 3: Using Augmentations to Simulate Dropout; Final Model Training

After the final data augmentation approach was selected, a data removal add-on approach is formulated, introducing random cutout techniques **[8]**. Specifically, this final augmentation is applied after the augmentations of the selected pipeline. This technique is performed by randomly erasing blocks of certain dimensions from an image, given the pipeline established in **Experiment 2**. Instead of giving this augmentation its own individual experiment, two final models (models that are trained on the entire training set given the augmentation pipeline chosen by **Experiment 2**) are trained. One of these models has the random cutout augmentation appended to the end of its respective augmentation pipeline, while the other's pipeline is unchanged from that chosen in **Experiment 2**. The final weights for the model are obtained by training on the full training set of 26,698 images extracted from ROI images (minus any images with 'unknown' classification and images placed in the validation set). *Keras callbacks* are used to obtain the weights with best validation accuracy over a period of 20 epochs.

Discussion and Results

The main intent of these experiments was to give an idea of what augmentations can help our medical classification models extrapolate to the world outside of testing data, while also obtaining good results on testing metrics. The first thought is to identify a variety of the most common data augmentation strategies and compare them individually in the domain of cancer classification. **Figure 1** gives a detailed look at examples of the augmentations selected on a subset of training images. The individual augmentation parameters are chosen in a way that follows a paradigm of only slight modification to the original image because 1) I wanted to keep relevant information that was encoded in the original image and 2) augmentations will be combined in the future and I didn't want alterations to accumulate to a level where we lose relevant information from the original image.

In my experiments I only train on a subset of the training data for two reasons. First, by taking only a subset of the training data, we can artificially increase the impact of data augmentations on model accuracy, allowing us to visualize which augmentations perform better. And second, a smaller training set makes training faster. **Figure 2** and **Figure 3** demonstrate the average validation accuracy and maximum validation accuracy along with the arguments used for each augmentation for the subsampled 1350 image set and 2700 image set respectively.

Augmentations are categorized into the *'safe'*, *'watchlist'*, and *'dangerous'* bins as described at the end of **Experiment 1**, depending on performance achieved. **Figure 4** shows which augmentations belong to each category after experiment completion. This visualization gives some insight into what types of augmentations help performance. Orientation approaches like rotations and mirror flips have great performance, which makes sense because we are giving our model "new" data with no structural changes. For instance, stroma should be classified as stroma

regardless of whether the image is rotated 180 degrees or not. The same may not necessarily be true for changing the color properties or using a translation technique on a stroma image. The noise-related techniques also end up being assigned to the safe category, which I interpreted to be because some of the lower-level features in the image are less relevant compared to the higher-level features, which is the main identifier for different classifications. For instance, blurring obfuscates finer details whereas the general edges of larger bodies are conserved. Gamma seems to be a dangerous operation when applied to these images. The reason for this isn't clear but it may say something about how relevant luminance is in classification in this domain in a way that directly changing brightness is not. I was very surprised to see shearing categorized into the watchlist, as it was my intuition that slight translation techniques would preserve important images features while giving our model "new" data, but more experimentation on this method, along with the other watchlist methods, is discussed in the next experiment.

The combination approach in **Experiment 2** assumes that the combination of all 'safe' augmentations yield a more robust and accurate classifier than any individual approach. There are two reasons for this assumption: 1) for the brevity of experimentation, as there are far too many combinations of all techniques considered, and 2) it seems to be a benign assumption, as the 'safe' techniques seem to retain relevant information when applied in unison. This assumption aside, the main focus of this experiment is on the combination of 'watchlist' categorized augmentations and seeing how alterations in the middle of the spectrum of effectiveness change the success of our model. **Figure 5** demonstrates the average validation accuracy and maximum validation accuracy along with the arguments used for each combination approach. Contrast tended to perform the best out of all the techniques, consistently ending up in

the top maximum validation accuracy. In the end, the contrast + brightness approach faired just as well in maximum validation accuracy as contrast + brightness + shearing, so the latter is used as it statistically performs the same even with one extra augmentation. **Figure 6** demonstrates the augmented images resulting from the combination pipelines. The intuition behind combinations of augmentations is the hope that we are left with a robust model at the end of training, similar to the intuition behind style transfer on medical images **[9].**

Dropout is a well-known regularization technique in areas of deep learning used to prevent overfitting. Usually this is achieved by randomly omitting particular neurons from updating their gradients during training. Dropout has been shown to prevent neural network units from co-adapting, leading to major improvements in model performance **[10]** **[11]**. Since changing the model *structure* in this way is outside the scope of this project, I wanted to find a way to "emulate" this technique through augmentation. Random erasing is the process of replacing some area of pixels in an image with some constant value **[8]**. The intuition is, if we can remove sections of the image that would otherwise be a clear indication of a certain classification, we can force the model to learn new features that become relevant in the absence of removed features. **Figure 7** demonstrates the random erasing technique on vanilla images.

We experiment on random erasing efficacy in the setting of two final models for two reasons. The first is the intuition that given more data, random erasing will have a greater ability to perform as an ad-hoc dropout substitute. For example, a small experimental training set could, by chance, have an overrepresentation of samples that are immune to the effects of random erasing (i.e., images with many instances of features that tip off the model towards the correct classification, without considering other features that may be indicative of the same classification). The last reason is simply due to the fact that we are close to the end of

experimentation in any case and judging the performance of two final models as opposed to one is hardly an inconvenience considering the benefit of results that are clear and do not need to be speculated on or extrapolated. **Figure 8** details the accuracy values of the final model. The random erase variant prevails with a maximum validation accuracy of 0.5% higher than the model without this augmentation applied. **Figure 9** demonstrates the images generated by the final image pipeline (that is, with random erasing).

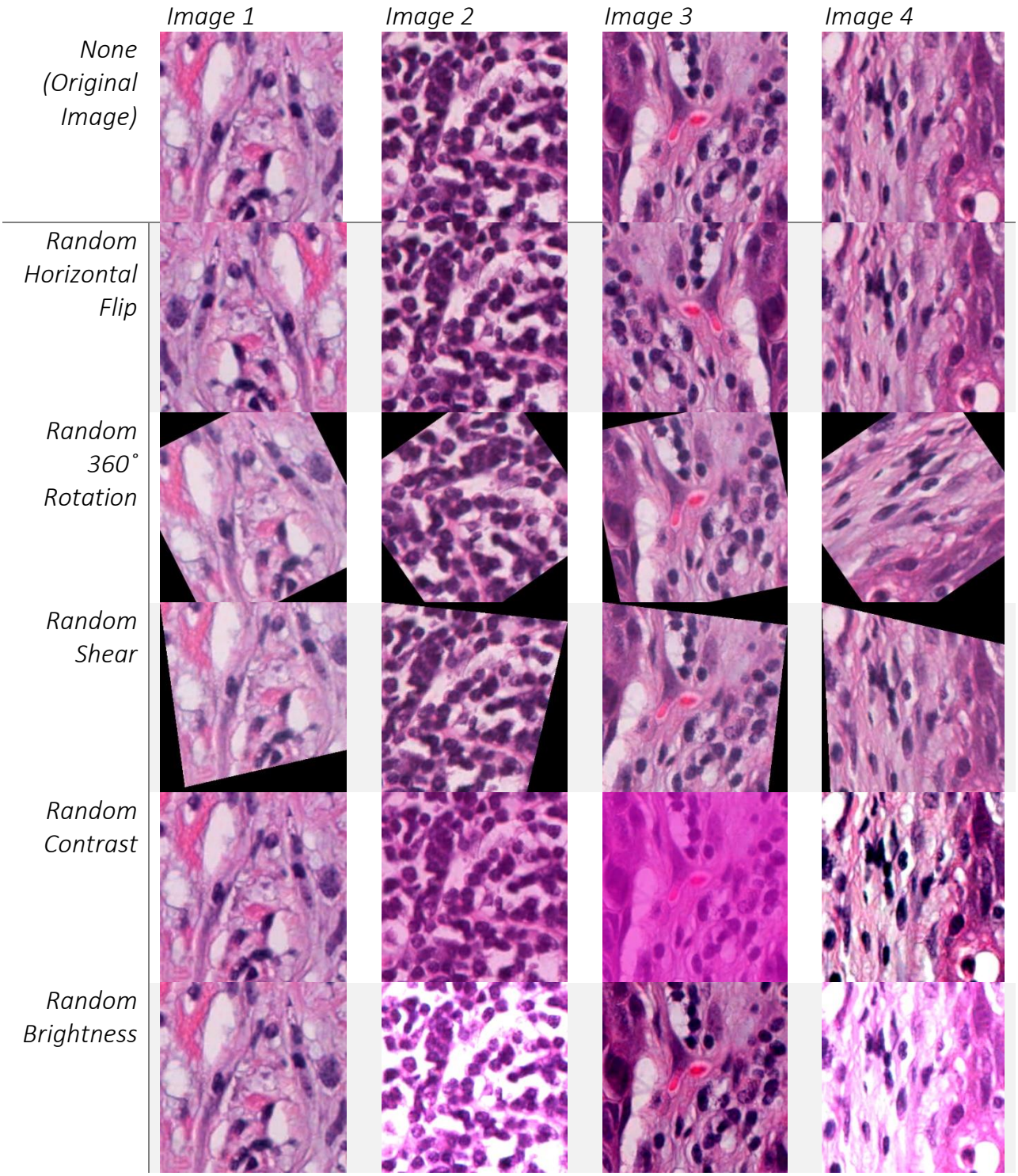
Our final model is tested on a final test set of 2510 images (2902 images minus 392 labeled as 'unknown') without augmentation and achieves a surprising test accuracy of 86.5% on the AUROC metric (there is a puzzling discrepancy between test and previous validation accuracy scores, this is considered and explained below). For experimental purposes, the other final model (trained on images without random erasing) is tested in the same manner and achieves a test accuracy of 84.74% on the AUROC metric. **Figure 10** showcases the final testing results for both models.

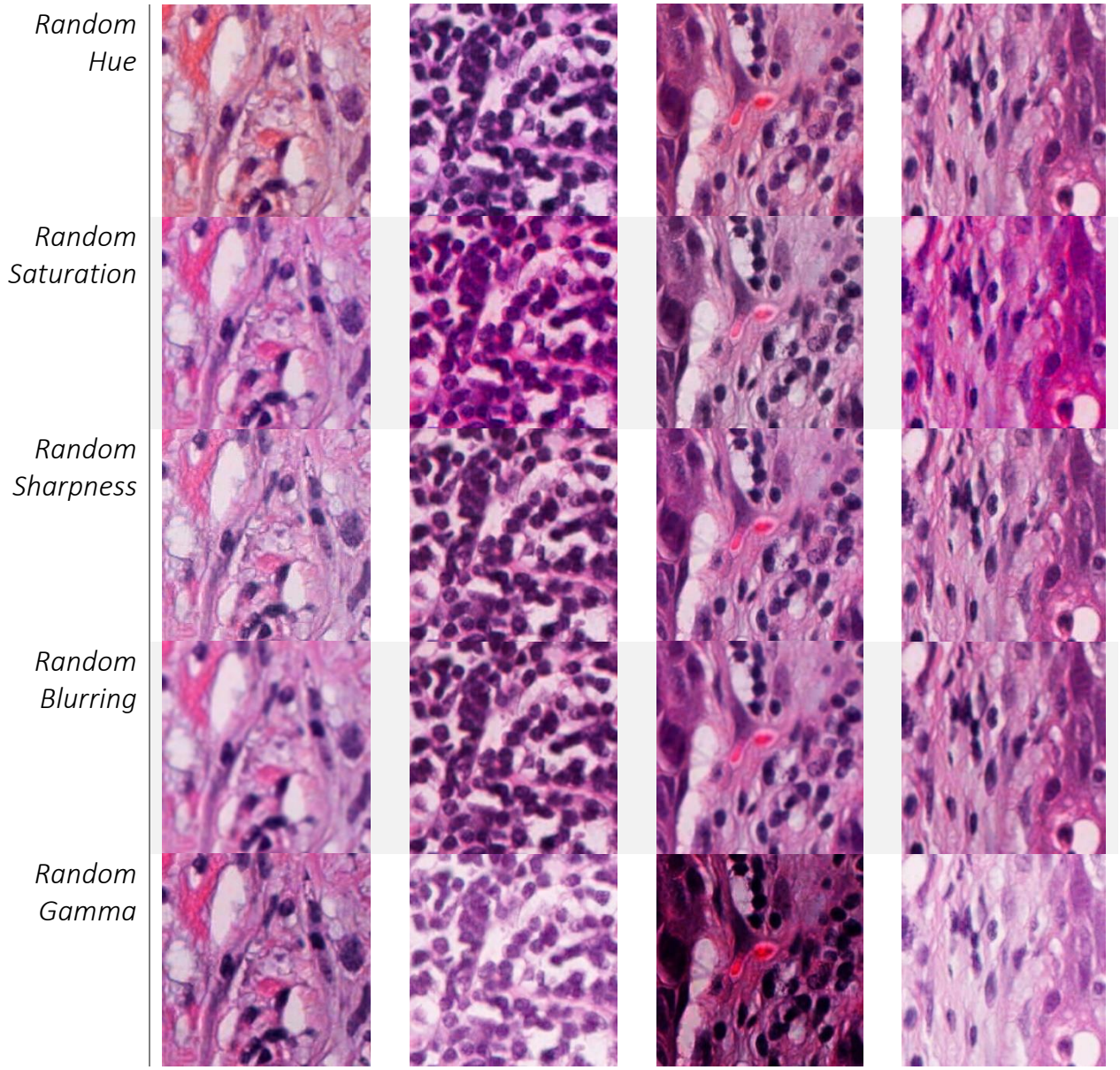
The goal of these experiments was to investigate what augmentations are beneficial to constructing a robust model in the domain of cancer classification. The results gathered by these experiments suggest that combinations of relevant augmentations, that may lead to a set of training images that tend toward "irrelevant" augmentations, can result in a robust model that still achieves relatively high accuracy. They also suggest that the counter-intuitive augmentation of information removal is an effective technique for increasing model performance. This being said, a *logistic error was discovered toward the end of experimentation*, where validation images were augmented as well. This led to validation scores suffering due to augmentations being applied to validation set images during the validation stages. In theory this shouldn't have major impact on picking the best augmentations (the validation accuracy instead reflects model

performance on altered images instead of unaltered ones, where altered images still retain relevant information for the classifier to work with), but given more time, these experiments should be replicated without this mishap. In hindsight, a larger subsampled set should have been used for experimentation, as by the end of this project I realized a 1350 image training set leads to variability in results and using a larger set of data for experimentation would have led to more reliable and concrete insights. Another possible limitation of my particular approach may be the assumption made in combining augmentations, and further work could be done on finding the best combination of many approaches, where one can rank them more effectively in this particular training domain. Furthermore, it is hard to quantify how robust the final model is given the data, which is a key question. Given more experience, some of which I have gained through this project, and more time, a more thorough approach could be pursued to definitively answer this question, along with the other concerns mentioned, moving forward.

Appendix

Figure 1: Examples of individual augmentations (applied to ROI images for demonstration purposes)





Source: Experiment 1

Figure 2: Individual image augmentation results (450 images per class; averaged over 2 runs; 30 epochs; sorted descending by maximum AUROC)

Augmentation	Arguments/Parameters	Average Validation AUC	Max Validation AUC
Random Brightness	Scale Factor Interval = (0 to 0.4)	0.7102	0.824
Random Horizontal Flip	None	0.712	0.819
Random Sharpness	Scale Factor Interval = (0 to 5)	0.717	0.815
Random Saturation	Scale Factor Interval = (0.75 to 1.5)	0.696	0.807
Random Hue	Scale Factor Interval = (0 to 0.06)	0.703	0.8
Random Blurring	Kernel Size Interval = (1 or 3 or 5 or 7) Sigma Size Interval = (1 to 2)	0.712	0.793
Random Double-Axis Shear	Shear Factor Interval = (-0.25 to 0.25) Fill Value = (0, 0, 0)	0.662	0.789
Random 360˚ Rotation	Degree = 360 Interpolation = Bilinear Fill Mode = Constant Fill Value = (0, 0, 0)	0.678	0.785
None (Original Image)	None	0.671	0.784
Random Contrast	Contrast Factor Interval = (0.75 to 1.5)	0.68	0.772
Random Gamma	Gamma Scale Factor Interval = (0 to 0.5) Gain Scale Factor Interval = (0 to 0.5)	0.57	0.635

Source: Experiment 1

Figure 3: Individual image augmentation results (900 images per class; 30 epochs; sorted descending by maximum AUROC)

Augmentation	Arguments/Parameters	Average Validation AUC	Max Validation AUC
Random Hue	Scale Factor Interval = (0 to 0.06)	0.761	0.838
Random Sharpness	Scale Factor Interval = (0 to 5)	0.741	0.824
Random Blurring	Kernel Size Interval = (1 or 3 or 5 or 7) Sigma Size Interval = (1 to 2)	0.721	0.817
Random Horizontal Flip	None	0.736	0.814
Random 360˚ Rotation	Degree = 360 Interpolation = Bilinear Fill Mode = Constant Fill Value = (0, 0, 0)	0.74	0.81
Random Saturation	Scale Factor Interval = (0.75 to 1.5)	0.715	0.802
None (Original Image)	None	0.74	0.802
Random Double-Axis Shear	Shear Factor Interval = (-0.25 to 0.25) Fill Value = (0, 0, 0)	0.73	0.794
Random Brightness	Scale Factor Interval = (0 to 0.4)	0.718	0.793
Random Contrast	Contrast Factor Interval = (0.75 to 1.5)	0.73	0.789
Random Gamma	Gamma Scale Factor Interval = (0 to 0.5) Gain Scale Factor Interval = (0 to 0.5)	0.5605	0.69

Source: Experiment 1

Figure 4: Augmentation techniques categorized based on performance in Experiment 1 (where *Safe* is beneficial, *Watchlist* requires more experimentation, and *Dangerous* is removed from experimentation)

Augmentation	Category
None (Original Image)	N/A
Random 360˚ Rotation	Safe
Random Blurring	Safe
Random Brightness	Watchlist
Random Contrast	Watchlist
Random Double-Axis Shear	Watchlist
Random Gamma	Dangerous
Random Horizontal Flip	Safe
Random Hue	Safe
Random Saturation	Safe
Random Sharpness	Safe

Source: Experiment 1

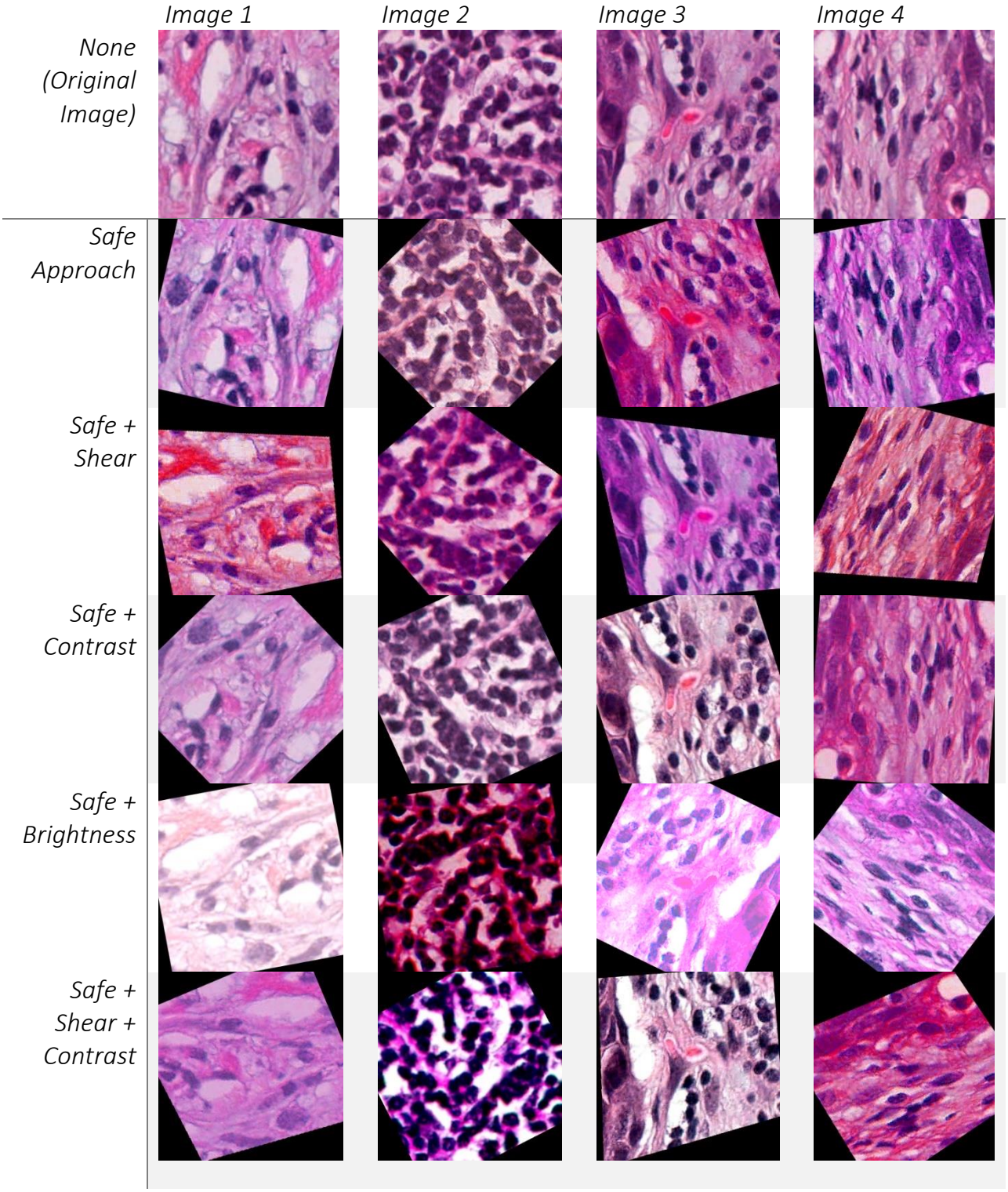
Figure 5: Accuracy results for combinations of augmentations (900 images per class; sorted by maximum AUROC)

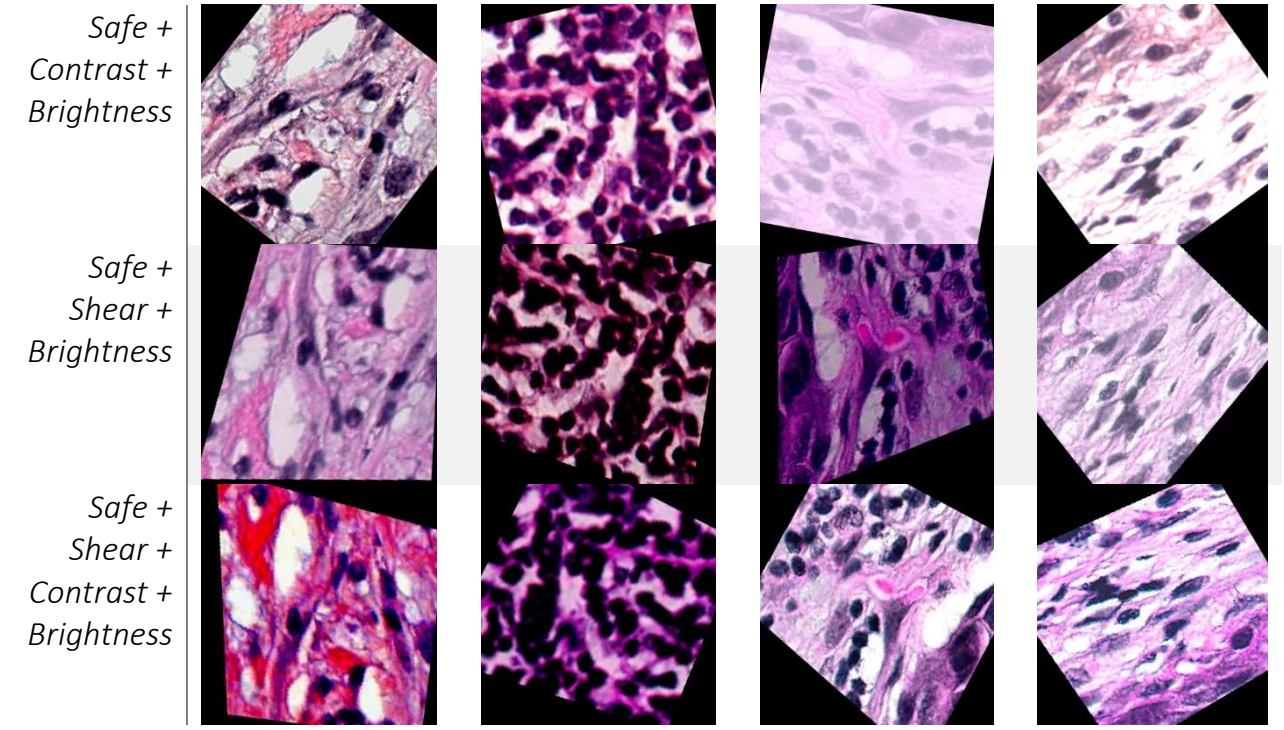
Note: There was a small confusion on my part, where validation data was augmented along with training data, causing many figures, like Figure 5, to read as if non-augmentation is superior, this is mentioned in the limitations of the experiments in the discussion. But overall, this should not affect the main point of the experiments.

Augmentation Approach	Average Validation AUC	Max Validation AUC
Original	0.739	0.819
Safe	0.733	0.786
Safe + Contrast + Brightness	0.695	0.776
Safe + Shear + Contrast + Brightness	0.68	0.776
Safe + Shear + Contrast	0.679	0.772
Safe + Brightness	0.721	0.762
Safe + Contrast	0.695	0.761
Safe + Shear	0.696	0.76
Safe + Shear + Brightness	0.676	0.73

Source: Experiment 2

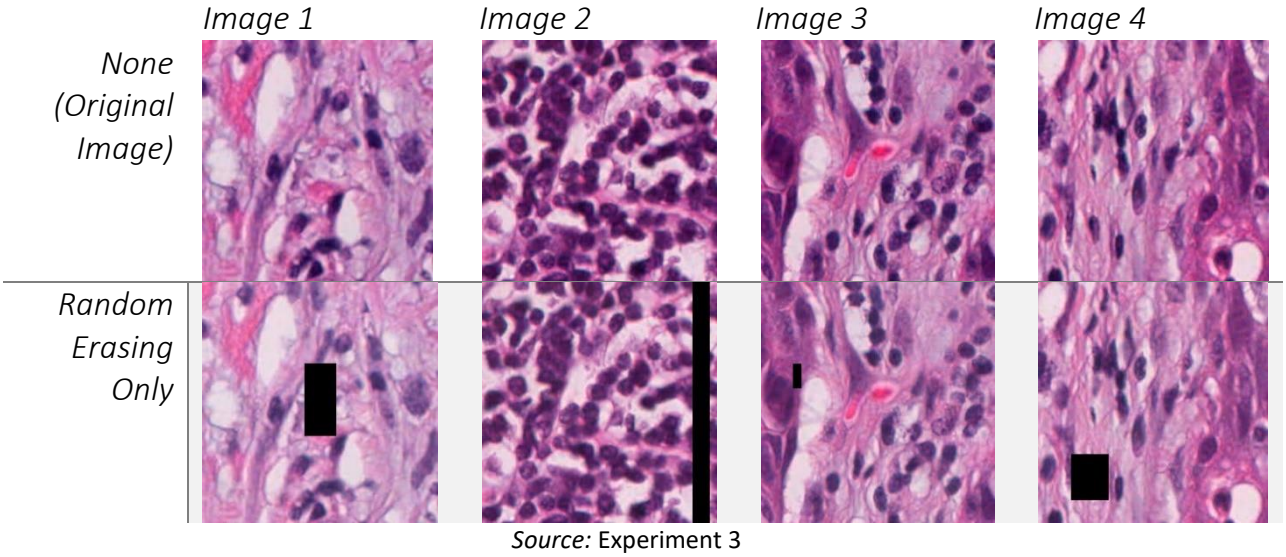
Figure 6: Combinations of augmentations example images (applied to ROI images for demonstration purposes)





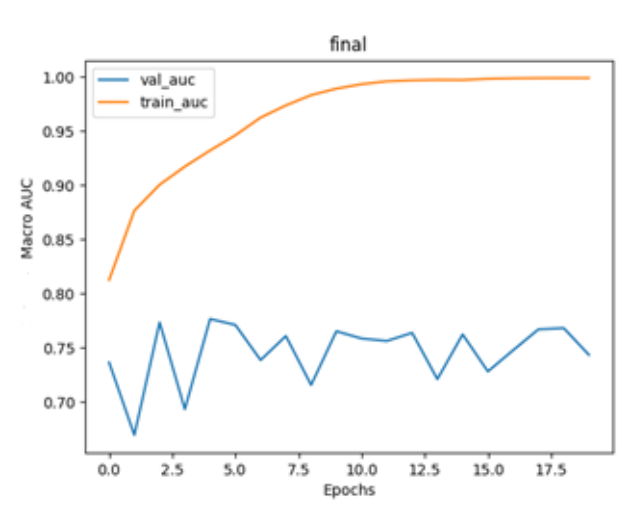
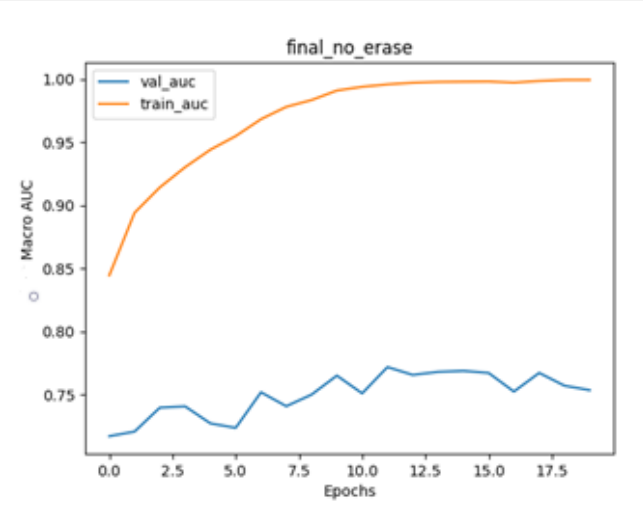
Source: Experiment 2

Figure 7: Random Cutout/Erasing Augmentation Examples (applied to ROI images for demonstration purposes)



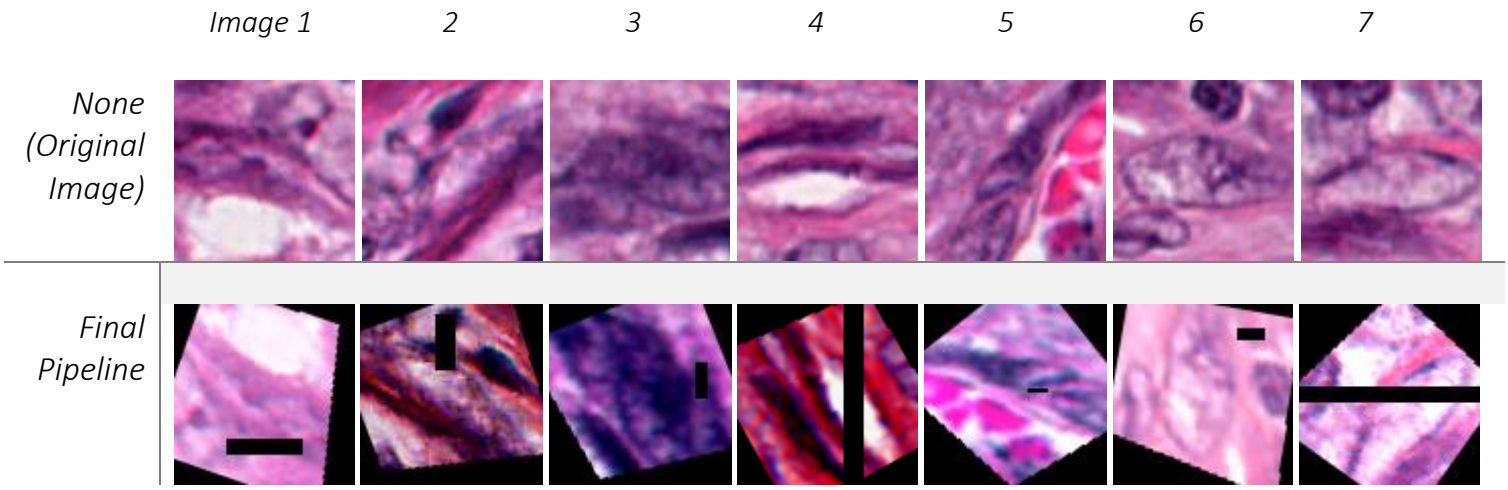
Source: Experiment 3

Figure 8: Final model results comparison (Full training dataset used; sorted descending by maximum AUROC)

Augmentation Approach	Training Figure	Average Validation AUC	Max Validation AUC
Random Erasing Applied		0.7459	0.7766
Without Random Erasing		0.7498	0.7717

Source: Experiment 3

Figure 9: Final augmentation pipeline examples (90x90 extracted cell images used)



Source: Experiment 3

Figure 10: Evaluation of final model with random erasing (top) and final model without random erasing (bottom) on the test set with a batch size of 48.

Note that accuracy scores here are much higher due to the metrics coming from a model that is being evaluated on non-augmented data. See limitations of the experiment for more details.

```
49/49 [==============================] - 13s 143ms/step - loss: 0.8076 - macro_auc: 0.8650
[0.8076194524765015, 0.8649534583091736]
```

```
53/53 [==============================] - 11s 150ms/step - loss: 1.6390 - macro_auc: 0.8474
[1.6389769315719604, 0.8473958969116211]
```

Source: Final Testing

Bibliography

[1] Shorten, C., Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. J Big Data 6, 60 (2019). https://doi.org/10.1186/s40537-019-0197-0

[2] https://www.tensorflow.org/api_docs/python/tf/image

[3] https://www.tensorflow.org/addons/api_docs/python/tfa/image

[4] https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/image

[5] https://matplotlib.org/3.4.3/contents.html

[6] https://arxiv.org/abs/2102.09099

[7] https://arxiv.org/abs/2006.04836

[8] https://arxiv.org/abs/1708.04896

[9] https://arxiv.org/abs/2102.01678

[10] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov; 15(56):1929−1958, 2014. https://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf

[11] Leibig, C., Allken, V., Ayhan, M.S. et al. Leveraging uncertainty information from deep neural networks for disease detection. Sci Rep 7, 17816 (2017). https://doi.org/10.1038/s41598-017-17876-z