Automated Classification of Breast Ultrasounds

Samuel Ivuerah [redacted]

1. Introduction

This project uses machine learning techniques to improve upon the metrics of the ResNet-18[2] network used in MedMNIST-v2[1]. The BreastMNIST dataset from MedMNIST-v2[1] resized from 500x500 grayscale images to 28x28, contains 780 samples split into training, validation, and testing images of breast ultrasounds with an approximate ratio of 7:1:2, respectively. BreastMNIST classifies the ultrasounds into two labels/classes, "malignant" or "normal, benign".

Goals:

- Develop a ResNet-18-based binary image classifier for the categories of BreastMNIST.
- Surpass the MedMNIST-v2's ResNET-18 metrics of AUC (>= 0.901) and ACC (>= 0.863) in the new image classifier.

2. Literature Review

ResNet models and other Convolutional Neural Networks (CNNs) are widely used to make deep neural networks, especially in medical imaging classification tasks such as BreastMNIST. For instance, the MedMNIST-v2[1] study demonstrated the effectiveness of ResNet-18[2] when it comes to classifying images on a dataset called PnemoniaMNIST, which happens to share similarities with our BreastMNIST dataset: Being of greyscale 28x28 images as a binary classification task. The ResNet-18 model in the study achieved an AUC of 0.956 and an ACC of 0.864, possibly implying that ResNet-18 is a good foundation for binary classification of medical images. Similarly, another study on Laparoscopic-Image-Distortion-Classification[3] using ResNet-18[2] achieved a training accuracy of 98.75% and a validation accuracy of 97.67%. Although the task was multi-classified, these high scores indicate that ResNet-18 is a robust model for medical image classification, especially when adapted to specific tasks.

3. Technical Details

The proposed model has been made by modifying the original ResNet-18 neural network: The first convolutional layer has been adjusted to accept single-channel greyscale images from the BreastMNIST dataset as opposed to the original convolutional layer designed to accept three-channel RGB images. The padding, kernel size, stride, and bias of the first convolutional layer are set to 1, 3, 1, and True, respectively, To preserve the spatial dimensions, maintain information at the borders of the images, enhance the capture of local features of the small images, increase overlapping of images to extract finer details of the small images and allow the activation functions within ResNet-18[2] to fit the dataset better. Succinctly, these changes to the first layer help extract more intricate details from the BreastMNIST dataset and make the model more flexible for this binary classification.

The fully connected layer has been replaced with a sequential layer of a dropout layer and a final linear layer. The dropout layer helps prevent overfitting, making it especially important for a small dataset such as BreastMNIST. The linear layer's outputs are modified to match BreastMNIST instead of the original 1000 classes from ResNet-18[2], for binary classification.

The Cross-Entropy Loss function is used during training as it is suitable for binary classification by computing log probabilities for the two classes in BreastMNIST and applying a softmax function to allow direct usage of the output scores from the model:

$$Softmax = f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}} \qquad CrossEntropyLossFunction = CE = -\sum_i^C t_i \cdot \log(f(s_i)_i)$$

S denotes the model's raw scores before the softmax activation, C and J are the total number of classes, T is the target label, and I denotes the $I - th$ class.

The Stochastic Gradient Descent (SGD) optimiser minimises the loss function, for a maximum of 30 epochs, in conjunction with a StepLR learning rate scheduler to help the optimiser converge faster, combined with a weight decay to help prevent overfitting. An early stopping strategy – evaluating on the validation set and monitoring the mean of AUC and ACC – is used to help prevent overfitting. A Google T4 GPU was used to conduct all training, validations, and tuning using the PyTorch framework, where a random search technique was used to conduct a hyperparameter search for parameters in Table 1.

4. Evaluation

The proposed model's metrics in Table 2 suggest that it outperforms MedMNIST-v2's ResNet-18[1]: AUC has improved by +0.056 (+6.2%), and ACC has improved by +0.06 (+7%), indicating that the proposed model has a better distinction between the classes on BreastMNIST and can accurately classify a higher proportion of images, satisfying both goals. Although, the proposed still seems to be overfitting, most likely related to the small dataset of BreastMNIST: From a production setting, the model may not as such.

However, comparing the metrics, (Average Val AUC: 0.890 Average, Val Accuracy: 0.865, Average Test AUC: 0.883 Average Test Accuracy: 0.846) of the new model with 5-Fold validation, shown in Table K, indicates that the model may be comparable to or worse than MedMNIST-v2's ResNet18[1] as the AUCs and ACCs are lower on 5-Fold validation. 5-fold validation metrics provide a more accurate representation of the new model's performance since the model is trained on folds over a merged BreastMNIST dataset.

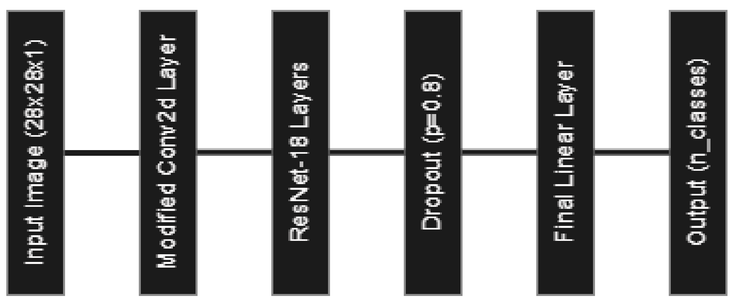
The difference between AUC and ACC on the test set for the new model is +0.034, (+3.68%); AUC is a metric computed by summation of all accuracies across all possible threshold values, whereas ACC is computed at a threshold of 0.5; It is probable the model's accuracy at the 0.5 threshold is lower than AUC, the model may perform better at other thresholds. Similarly, there is a difference between the AUPR and F1 score on the test set: AUPR considers the above metrics for different thresholds, whereas the F1 score is a harmonic mean of precision and recall at a specific threshold; thus, the F1 score can and is lower than the AUPR metric.

Methods	Train Split						Test Split					
	AUC	ACC	AUPRC	Recall	F1	Precision	AUC	ACC	AUPRC	Recall	F1	Precision
MedMNISTv2's ResNet-18	-	-	-	-	-	-	0.901	0.863	-	-	-	-
Modified ResNet18	**0.982**	**0.949**	0.979	0.933	0.935	0.936	**0.957**	**0.923**	0.947	0.910	0.904	0.898

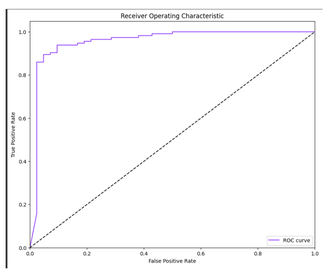
Table 2

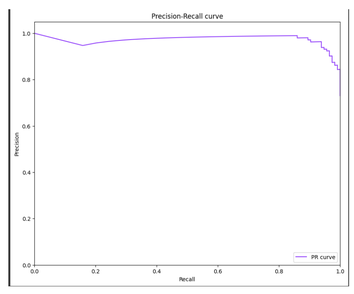
Learning Rate	Mini-batch Size	Dropout	Momentum	Weight Decay	Step Size	Gamma
0.001	32	0.8	0.9	0.0001	20	0.7

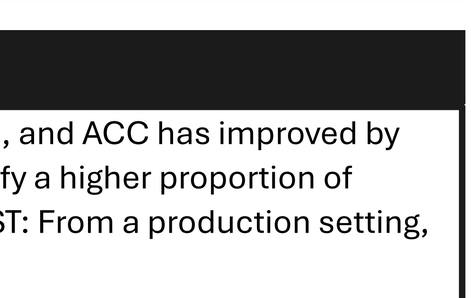
Table 1



Modified ResNet-18 Diagram (Made using Diagrams.net)







References:

[1] Yang, J. et al. (2023a) MedMNIST v2 - a large-scale lightweight benchmark for 2D and 3D Biomedical Image Classification, Nature News. Available at: https://www.nature.com/articles/s41597-022-01721-8 (Accessed: 25 April 2024).

[2] ResNet18 - He, K. et al. (2015) Deep residual learning for image recognition, arXiv.org. Available at: https://doi.org/10.48550/arXiv.1512.03385 (Accessed: 25 April 2024)

[3] Belmokeddem, M., Khemis, K. and Loudjedi, S. (2023) *Residual network (resnet-18) for Laparoscopic Image Distortion Classification*, *EasyChair Home Page*. Available at: https://easychair.org/publications/preprint/LBxp (Accessed: 25 April 2024).