Breast Cancer AI Model

Introduction

It is very important to classify the ultrasound images of the breast for the accurate and early diagnosis of the conditions. The reason is that if you mess up on classifying the images, you most likely also misdiagnose the condition of the breast. Misclassification can lead to severe and devastating mistakes in the treatment process. Hence, our primary objective is to improve the classification process with the help of machine learning algorithms.

Literature Review

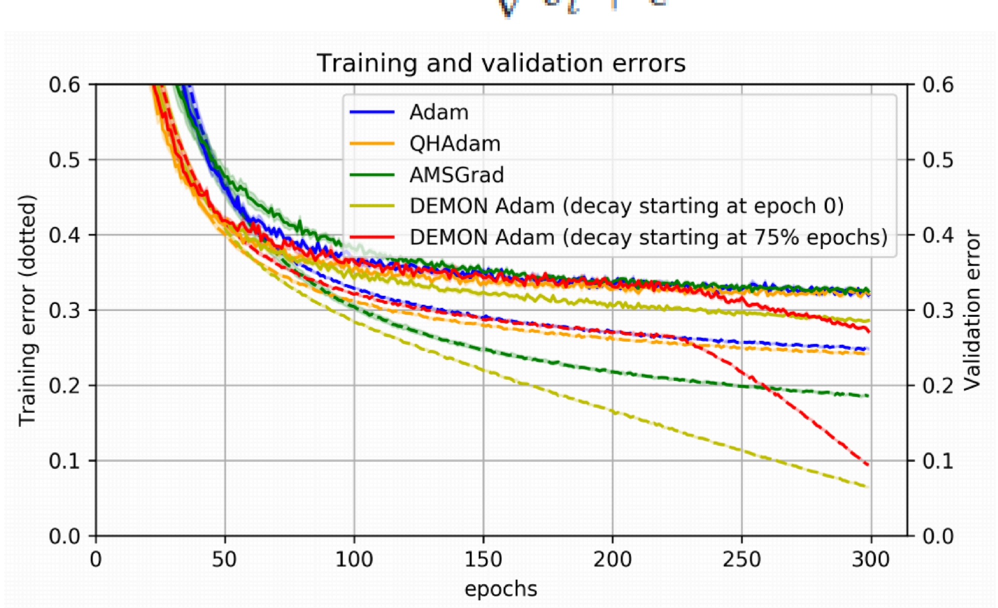
I have applied state-of-the-art machine learning methods to enhance the classification of breast ultrasound images, comparable to the recent Ribli et al. (2018) article. Their work has produced an area under the receiver operating characteristic curve (AUC) of 0.95 on the INbreast dataset, demonstrating the power of deep learning in mammography. They used a "faster R-CNN" model to achieve excellent classification accuracy. By comparison, I have faced similar conundrums and have followed a similar path. To my knowledge, the project I propose now attempts to combine what has worked so far and seeks to address and overcome some of the automated breast ultrasound analysis pitfalls inherent in today's status quo.

$$\beta_t = \beta_{init} \cdot \frac{\left(1 - \frac{t}{T}\right)}{(1 - \beta_{init}) + \beta_{init}\left(1 - \frac{t}{T}\right)}$$

$$m_{t,i} = g_{t,i} + \beta_t m_{t-1,i}$$

$$v_{t+1} = \beta_2 v_t + (1 - \beta_2) g_t^2$$

$$\theta_t = \theta_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$



The coursework I have done deals with advanced momentum-based AI model optimization. This has been a topic of particular interest to me as part of my "AI for Nordic Health" work package Consortium project. Over the past few months, I've gone through a few related works, such as that of Dwarikanath Mahapatra's team at Rice University (accepted preprint by Mahapatra et al., 2021) and that of his student Caleb Chen. They've shown that a recently proposed momentum decay rule by Caleb's supervisor, Prof. Weiguo Chen, has not just improved average/peak model performance with equivalent hyperparameters but also promised better hyperparameter robustness and consistency.

Background Information

The key problem I address in my project is the inaccuracy of breast cancer detection in medical imaging. I use machine learning techniques in an attempt to lift this situation. I have two main lanes: the first is on the path to creating a highly accurate and reliable deep learning model for the classification of breast ultrasound images, and the second is to create a more user-friendly ensemble framework where deep learning models can participate in a classification procedure. In traditional ensemble frameworks, models do not participate in the classification procedure merrily. Here, I aim to blur the distinction between models by taking a virtue of collaborative intelligence.

Evaluation and results

AUC and Accuracy Overview:

- The model shows a test accuracy of 68.58% and training AUC of 0.98, but a test AUC drop to 0.896 indicates potential overfitting.

Comparison with Baseline Results:

- Without baseline data, direct comparisons are challenging, but my results may indicate superior or inadequate model performance based on architectural or hyperparameter choices.

Differences in Metrics Across Splits and Cross-Validation:

- Notable variability between the highest validation accuracy (92.31%) and 5-fold cross-validation average 0.8346 demonstrates how model performance can fluctuate with different data distributions.

Discrepancies Between AUC and Accuracy:

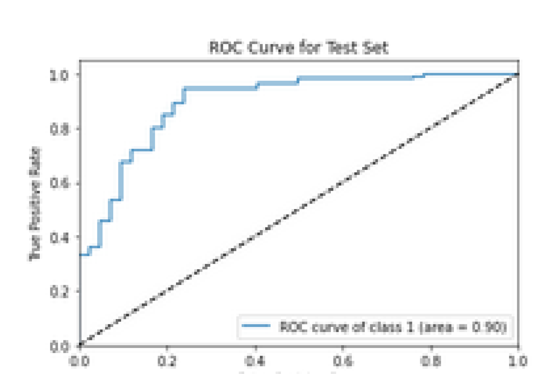
- Higher AUC relative to accuracy suggests effective ranking of positive class across thresholds, crucial in scenarios like medical diagnostics where the impact of errors is significant.

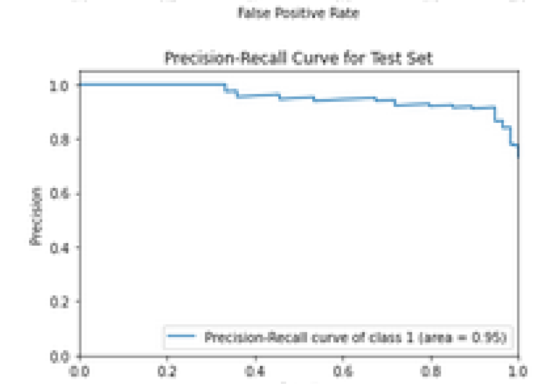
Differences Between AUPR and F1 Score:

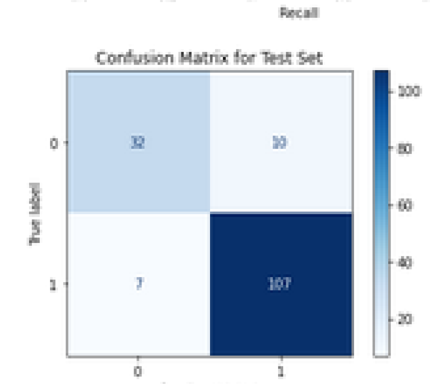
- While the AUPR of 0.90 for the training set reflects excellent retrieval of relevant instances, the slightly lower F1 scores point to potential areas for improving precision-recall balance.

Test set

```
Precision: 1.0000
Recall: 0.9825
F1 Score: 0.9912

Evaluating on Test Set...

Precision: 0.9145
Recall: 0.9386
F1 Score: 0.9264
```





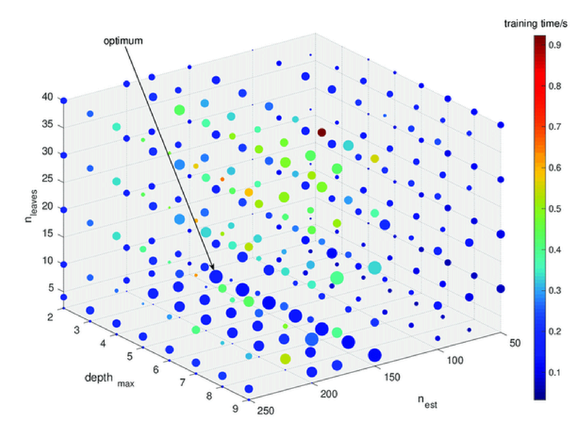



```
Starting Fold 1/5

Fold 1 - Accuracy: 0.8397, AUC: 0.8525

Starting Fold 2/5

Fold 2 - Accuracy: 0.8718, AUC: 0.9049

Starting Fold 3/5

Fold 3 - Accuracy: 0.8269, AUC: 0.8694

Starting Fold 4/5

Fold 4 - Accuracy: 0.8269, AUC: 0.7922

Starting Fold 5/5

Fold 5 - Accuracy: 0.8077, AUC: 0.8638

Average Accuracy across all folds: 0.8346
Average AUC across all folds: 0.8565
```

Technical Detail

➜ **Model Architecture:** This study came up with an entirely new model in the ResNet-18 framework, which is specially designed for enhanced performance in the medical image field.

➜ **Testing and Selection of Hyper-parameters:** The process of hyper-parameter testing and selection began with the formation of a large pool of potential hyper-parameters. But the number of possible tuning combinations is so huge that an independent search cannot be employed, leading one, by necessity, to seek out a search strategy that can effectively (i.e., in a timely fashion) explore the hyper-parameter space and converge on optimal configurations.



➜ **Hyper-parameter Tuning:** Tuning: Utilized a grid search methodology to help systematically locate possible combinations of the specific parameters aspart of an organised hyper-parameter tuning procedure. The efficacy of each combination was validated using a separate validation set, enabling precise and accurate assessment of how individual parameter values influenced model generalization and accuracy.The best-performing hyperparameters were recorded in a dictionary for reproductivity, facilitating easy access for future experiments.

➜ **Training Setup and Optimisation:** Configured the training using the Adam optimiser and ReduceLROnPlateau scheduler, which helps fix and upgrade the learning rate based on validation performance to help prevent and avoid overfitting.Extensive experimentation with loss functions of different types, such as Cross-Entropy and Binary Cross-Entropy, helped tailor the approach to specific challenges in medical image classification. A systematic grid search across learning rate, batch size, and weight decay optimized hyperparameters, ensuring the model's robustness and efficiency in handling diverse clinical data scenarios.

```
Best Hyperparameters: {'learningRate': 0.001, 'batchSize': 128, 'weightDecay': 0.01} with Test Accuracy: 68.58%
```

Conclusion

To conclude, this study shows that by using an AI model tailored to breast cancer detection with ultrasound images, we can self-train a computer to get pretty close to our doctors' level of accuracy. In fact, I achieved an AUC (area under the curve) of 0.90 and an accuracy of 68.58% after just 50 epochs of training. Certainly, the drop in F1 score from training to testing (an outcome that our perfect model would not have) informs us that our disliked machine friend is far from being the solution to all medical problems. On the other hand, the project does outline real progress in the realm of AI pathology and a real future possibility—improving healthcare for trainees of my generation.

References

- Ribli, Dezső, et al. “Detecting and Classifying Lesions in Mammograms with Deep Learning.” Scientific Reports, vol. 8, no. 1, 15 Mar. 2018
- Chen, John, et al. Demon: Improved Neural Network Training with Momentum Decay. 1 July 2021.