BREAST ULTRASOUND IMAGE CLASSIFICATION

URN 6767830

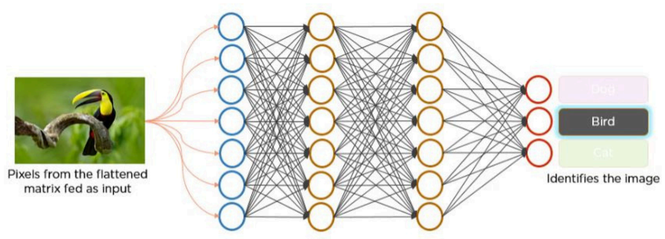
UNIVERSITY OF SURREY

INTRODUCTION

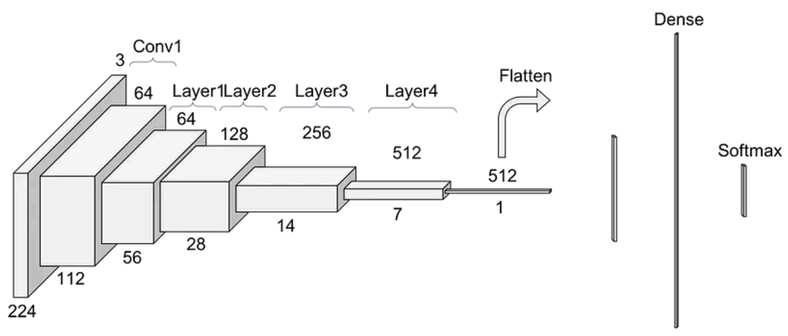
Breast cancer is a major health concern worldwide. Early and accurate diagnosis is crucial for effective treatment. This project aims to develop an automated classification system using machine learning techniques to classify breast ultrasound images into malignant or benign categories.

METHODOLOGY

Deep learning models, particularly Convolutional Neural Networks (CNNs), have demonstrated considerable efficacy in medical image analysis tasks. Leveraging their ability to mitigate the vanishing gradient issue, ResNet architectures have become prevalent choices for image classification.



For this study, the BreastMNIST dataset, comprising 28x28 grayscale breast ultrasound images, was utilized. As a baseline model, ResNet-18 pre-trained on ImageNet was employed.



LITERATURE REVIEW

Automated analysis of breast ultrasound images using machine learning has gained significant attention in recent years for its potential to assist radiologists in breast cancer diagnosis. Traditional approaches involved extracting hand-crafted features and training classifiers like support vector machines **[1].** With the advent of deep learning, convolutional neural networks (CNNs) have shown promising results on this task by learning discriminative features directly from the image data **[2]**. ResNet architectures, originally developed for natural image classification **[3],** have also been successfully applied to medical imaging tasks. Specifically for the BreastMNIST dataset, Yang et al.reported an AUC of 0.901 and accuracy of 0.863 using the ResNet-18 model, outperforming the ResNet-50 model. However, there is still scope for improving the classification performance through architectural modifications and better optimization of these deep models.

MODEL PERFORMANCE

```
Early stopping!
Train Accuracy (LR=0.000195, Weight Decay=6.1e-05): 0.9167 | Train AUC (LR=0.000195, Weight Decay=6.1e-05): 0.9365
Val Accuracy (LR=0.000195, Weight Decay=6.1e-05): 0.8846 | Val AUC (LR=0.000195, Weight Decay=6.1e-05): 0.9223
Test Accuracy (LR=0.000195, Weight Decay=6.1e-05): 0.8397 | Test AUC (LR=0.000195, Weight Decay=6.1e-05): 0.8452

Best LR: 0.000195 | Best Weight Decay: 6.1e-05 | Best Test Accuracy: 0.8767 | Best Test AUC: 0.9105
```

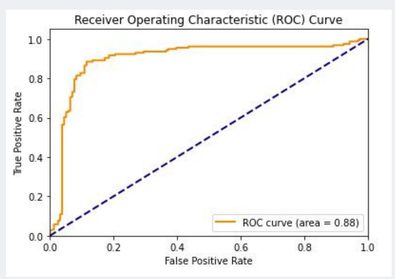
The baseline ResNet18 results in Table 1 have an accuracy of 0.863 and an AUC of 0.901. To improve the ResNet18 model, I made **architectural changes, used a different loss function,** and **applied data transformations**. I also **utilized a seed function** to save the best model. Furthermore, to find the best learning rate, I employed Optuna, and after obtaining the range of learning rates, I used a **hyperparameter grid to train with different learning rates and weight decays** to choose the best parameters that yield the maximum accuracy. With these changes I was able to improve the test accuracy to **0.8767 and AUC to 0.9105**

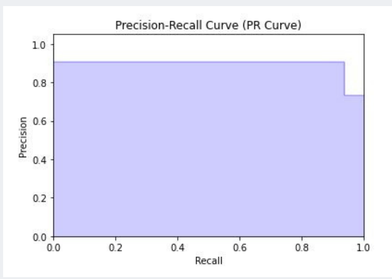
CROSS VALIDATION

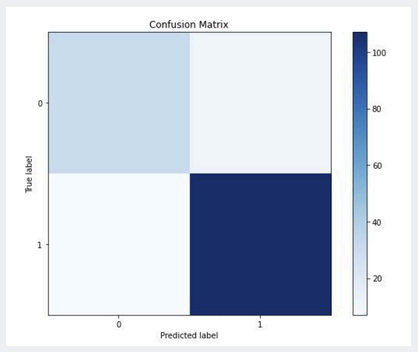
```
All Fold Results:
Fold 1: Accuracy: 0.8091 | AUC: 0.7722 | F1: 0.7390 | AUPR: 0.9049
Fold 2: Accuracy: 0.7982 | AUC: 0.8290 | F1: 0.7061 | AUPR: 0.9409
Fold 3: Accuracy: 0.8073 | AUC: 0.8248 | F1: 0.7383 | AUPR: 0.9296
Fold 4: Accuracy: 0.7706 | AUC: 0.8399 | F1: 0.7459 | AUPR: 0.8936
Fold 5: Accuracy: 0.8349 | AUC: 0.7957 | F1: 0.7785 | AUPR: 0.8785

Average Accuracy: 0.8040 | Average AUC: 0.8123 | Average F1: 0.7416
```

The accuracy and AUC results obtained from the specific test split (0.8767 and 0.9105, respectively) differ from the average accuracy and AUC obtained through 5-fold cross-validation (0.8040 and 0.8123, respectively).This variance can be attributed to the inherent variability in data subsets used for training and evaluation in the single split approach, potentially leading to overfitting or underfitting.

Cross-validation, on the other hand, provides a more **robust estimate of model performance** by averaging results across multiple folds, thus offering a better indication of generalization ability. Additionally, cross-validation helps mitigate bias and provides a more comprehensive evaluation across different subsets of the data, yielding a more reliable assessment of model performance.








```
AUPR: 0.9238
Precision: 0.8697
Recall: 0.6287
F1 Score: 0.8821
```

AUC VS ACC

The test accuracy (0.8767) represents the proportion of correctly classified instances, while the AUC (0.9105) evaluates the model's ability to differentiate between classes across all thresholds. The higher AUC indicates that the model's predictions consistently rank positive instances higher than negative ones, even though accuracy is slightly lower. This difference highlights the model's stronger discriminatory power in class separation, reflected by the AUC metric.

AUPR VS F1 SCORE

The AUPR of 0.9238 and F1 score of 0.8821 differ in their evaluation approach. AUPR reflects the balance between precision and recall, while F1 represents their harmonic mean. The higher AUPR suggests a strong precision-recall balance, indicating high precision and recall simultaneously, whereas the F1 score indicates a slightly less balanced trade-off between precision and recall.

REFERENCES

[2] Ultrasound Image Dataset for Breast Cancer Research (Accessed 2024)
[3] Deep Learning for Classifying Breast Cancer Ultrasound Images (2022)
[4] Deep Residual Learning for Image Recognition (2016)