Report: Breast Cancer Classification using Machine Learning

Introduction

Breast cancer is a critical medical condition that requires accurate and efficient classification to aid in early detection and treatment. This project utilizes machine learning techniques to classify breast cancer ultrasound images into categories: normal, benign, and malignant. The classification is achieved by leveraging transfer learning and training deep learning models, namely ResNet-based architectures, on a publicly available dataset.

Dataset

The dataset used is **Breast Ultrasound Images Dataset** obtained from Kaggle. It consists of three categories of images:

1. **Normal**
2. **Benign**
3. **Malignant**

Dataset Statistics:

- **Total Images:** 1,578
- **Category Distribution:**
    - Normal: 266 images
    - Benign: 421 images
    - Malignant: 891 images

The images vary in dimensions, with a minimum resolution of 190×310190 \times 310190×310, a maximum of 1048×7191048 \times 7191048×719, and an average size of 616×501616 \times 501616×501.

Data Preprocessing

1. **Image Validation:** Skipped invalid or corrupted images.
2. **Augmentation (Training):**
    - Resize images to 128×128128 \times 128128×128
    - Random horizontal flip
    - Random rotations
    - Color jitter
3. **Normalization:** Normalize pixel values using mean and standard deviation values derived from ImageNet.
4. **Train-Test Split:** The dataset is split into 80% training and 20% testing.

Models Developed

1. **Baseline Model: ResNet50**
    - Pre-trained on ImageNet.

- Fine-tuned by replacing the classification layer with a new fully connected (FC) layer for the three classes.
- Additional FC layers with ReLU activation and dropout for better generalization.

2. **Enhanced Model**
   - Built on ResNet18 architecture.
   - Custom FC layers added, including Flatten, Dense, ReLU, and Dropout layers.

Training Process

- **Loss Function:** CrossEntropyLoss with class weights to handle class imbalance.
- **Optimizer:** Adam optimizer with a learning rate of 0.001.
- **Learning Rate Scheduler:** StepLR with a step size of 5 and gamma of 0.1.
- **Epochs:** 15
- **Batch Size:** 32

Model Evaluation Metrics

1. **Accuracy**
2. **F1-Score**
3. **Precision**
4. **Recall**
5. **Validation Loss**

Results

1. **Baseline Model Results:**
   - **Accuracy:** 87.97%
   - **F1-Score:** 87.96%
   - **Precision:** 88.08%
   - **Recall:** 87.97%
2. **Enhanced Model Results:**
   - **Accuracy:** 87.65%
   - **F1-Score:** 87.65%
   - **Precision:** 87.74%
   - **Recall:** 87.65%
3. **Final Model with Additional Data:**
   - **Accuracy:** 89.87%
   - **F1-Score:** 89.83%
   - **Precision:** 90.07%

- **Recall:** 89.87%

The addition of new data and model enhancements showed noticeable improvements in performance metrics.

Deployment

A user-friendly interface was developed using **Streamlit** to upload breast ultrasound images for classification. The application:

1. Preprocesses the uploaded image using the same normalization pipeline.
2. Predicts the class using the trained enhanced ResNet18 model.
3. Displays the prediction along with the uploaded image.

Conclusion

The project successfully demonstrates the use of machine learning in medical image classification, achieving high accuracy and reliable results. The deployment of a web-based interface further provides an accessible tool for breast cancer diagnosis.