Mariana Faria Pereira
Mestrado em Engenharia e Ciência de Dados; Licenciatura em Engenharia Química
https://github.com/mfaria-p/BREASTMNIST_CNN.git

JuniorAI

AI Hub Challenge: Beast Tissue Classifier

Project Overview

- Implemented an end-to-end CNN model for **BreastMNIST** (binary classification: malignant vs benign).
- Built pipeline: dataset loading → preprocessing → CNN training → evaluation → deployment with Gradio UI.

Dataset insights

- Used **BreastMNIST** (780 breast ultrasound images, resized to 28×28 grayscale).
- Task is **binary classification**:
  - 0 = malignant
  - 1 = normal + benign (two categories merged into one positive class).
- Small dataset size → risk of overfitting, harder to generalize.

Modeling

- First time implementing a **Convolutional Neural Network (CNN)** (previously only worked with ANNs).
- Designed a small CNN with **3 convolutional blocks + classifier**.
- Learned how CNNs handle **spatial information** in images (filters, pooling, feature maps).

Challenges

- **Understanding image preprocessing**: converting ultrasound images into 28×28 grayscale tensors suitable for CNN input.
- **Label confusion**: dataset documentation originally mentioned 3 categories (normal, benign, malignant), but BreastMNIST simplifies it into **2 classes** (malignant vs normal+benign).
- **Result display in Gradio**: outputs did not show exactly as expected
- Limited dataset → accuracy fluctuated, validation/testing needed careful interpretation.

Solutions

- Carefully checked **dataset metadata** (INFO dictionary) to confirm true label mapping.
- Normalized and reshaped images manually to match training pipeline.
- Monitored training with validation accuracy, saved the **best model checkpoint**.

Results

- Achieved reasonable performance (~70–80%+ test accuracy depending on runs).
- Visualized **confusion matrix** and classification report for deeper evaluation.

New skills acquired

- Hands-on experience building and training a CNN from scratch.
- Learned practical **image preprocessing** and how CNNs differ from ANNs.
- Gained familiarity with **MedMNIST datasets**.
- Built an **interactive Gradio demo**, connecting ML models with a simple UI.
- Understood the importance of **dataset clarity** and double-checking label mappings.

Takeaways

- CNNs are powerful for image tasks but require careful data preparation.
- Small research datasets are useful for prototyping, but real-world use demands larger, diverse datasets and validation.
- Combining coding (PyTorch) with deployment tools (Gradio) is essential for end-to-end ML skills.