A Deep-Learning Approach to Breast Cancer Screening from Mammography Images

Laurent Lejeune

June 7, 2024

Contents

1 Mammography Dataset

We use the publicly available Chinese Mammography Database (CMMD) [1], which originally contains $\approx$ 1871 patients screened for breast cancer. and apply relevant filtering criteria to remove:

- Patients with history of previous breast biopsy within 1 week, or any therapy for breast lesions prior to mammography
- Patients with breasts prosthesis
- Images with substantial motion artifact

Each patient is then diagnosed by an expert and assigned the following target variables:

- $y_p \in \{\text{benign}, \text{malignant}\}$ indicates the type of tumor.
- $y_a \in \{\text{calcification}, \text{mass}, \text{both}\}$ indicates the type of abnormality, where `both` means that both `calcification` and `mass` are present.
- $y_s \in \{\text{luminal-A}, \text{luminal-B}, \text{HER2-positive}, \text{triple-negative}, \text{missing}\}$ a subtype information (possibly missing).

2 Dataset exploration

References

[1] Hongmin Cai et al. “An online mammography database with biopsy confirmed types”. In: *Scientific Data* 10.1 (2023), p. 123.