

Product Requirements

Data Elements

Type	Fields
Images	Ultrasound frame (JPG/PNG, resized 224×224, normalized)
Labels	Benign (0), Malignant (1) from MMOTU class mapping
Biomarkers	Age (28–78), CA-125 (5–35 benign; 35–600 malignant, skewed), BRCA (0/1)
Metadata	Image ID, (optional) case ID for grouping/splitting

Collection & Storage

- Source images from MMOTU public dataset; synthetic biomarkers generated with fixed seed.
- Store in project directory with CSV manifest aligned to image IDs.

Preprocessing

- Resize to 224×224, RGB conversion, [0,1] normalization.
- Standardize tabular features (z-score) prior to model training; keep BRCA binary.
- Ensure case-level consistency for train/val/test splits when possible.

Anonymization & Privacy

- Use only de-identified public images; no PHI is collected or stored.
- Synthetic biomarkers are generated and not patient-derived.

Bias Mitigation

- Report per-class metrics; use class weights for imbalance.
- Regularize tabular encoder to avoid over-reliance on CA-125 cutoff.
- Document dataset limitations and synthetic feature constraints.

Compliance

- Research/education use only; not for clinical use.
- Respect dataset licenses/citations; no attempt to re-identify subjects.

Governance & Reproducibility

- Fix random seeds; version training code and model artifacts.
- Keep a run log (config, metrics, commit hash).