Project G5: Melanoma Detection using Knowledge Distillation for Mobile Phones

- **Validate teacher complexity**: ResNet-34 vs ResNet-50 vs EfficientNet-B2 on validation accuracy first.
- **Simplify hyperparameter search**: Proposed grid (T=[1,2], α=[0.1,0.5,0.7,0.9], LR=[0.00001–0.01], batch=[16–256], dropout=[0.1–0.5]) = hundreds of configs—**use Bayesian optimization we discussed in class** or prioritize T and α only.
- KD is appropriate for mobile deployment, but **establish teacher performance ceiling** (train ResNet50, measure AUC/ECE) before distillation—if teacher ECE>0.1, fix calibration first.

Actions before Milestone-2

- **Dataset splits**: Download HAM10000, create 70/15/15 train/val/test (stratified by melanoma label), **not 80/20 train/test with K-fold CV**—CV is too expensive for KD experiments.
- **Simple baselines**: Train logistic regression on metadata (age, sex, lesion location) + handcrafted features (color variance, asymmetry), expect ∼70% AUC baseline.
- **Train teacher first**: ResNet50 (ImageNet pretrained, frozen backbone, fine-tune head) with weighted BCE (weight=8 for melanoma), train >20 epochs, log AUC, ECE, and precision@95% sensitivity on val.
- **Temperature scaling**: If teacher ECE>0.05, apply post-hoc temperature scaling on validation set before KD—properly calibrated teacher is essential for distillation.
- **Recommended tracking setup**: W&B with fixed seeds, log loss curves (KL + BCE components separately), model size (MB at fp32/fp16/int8), inference latency (ms on CPU), memory footprint (MB RAM).

Ablations

- **Temperature T**: Compare T=1 (hard targets) vs T=2 (softer targets)—hypothesis is T=2 improves student calibration (lower ECE by 0.02–0.03) following Islam et al.
- **Loss weighting** α: Test α=0.5 (balanced) vs α=0.9 (teacher-heavy)—expect α=0.5 yields best AUC-ECE tradeoff, avoiding over-reliance on uncalibrated teacher.
- **Quantization**: Post-training quantization (fp32→int8) should drop AUC by <2% and ECE by <0.03—if degradation higher, try quantization-aware training (QAT).

Risks & mitigations

- **Risk**: HAM10000 has 11% melanoma imbalance, student may underfit minority class—**Mitigation**: Use weighted BCE (weight=8 for mel), focal loss (γ=2), or oversample melanoma 2:1 during training.
- **Risk**: Extensive hyperparameter grid (T, α, LR, batch, dropout, quantization) is computationally prohibitive—**Mitigation**: Prioritize T and α (4 configs), fix other hyperparams to literature defaults (LR=0.001, batch=64, dropout=0.2).
- **Risk**: Quantized student loses calibration (ECE spikes)—**Mitigation**: Measure ECE on quantized model separately, retrain with QAT if post-training quantization ECE>0.1.

Open questions

- How will you handle domain shift from HAM10000 (dermoscopy, clinical lighting) to user-taken smartphone photos (variable lighting, distance, angle)—will you test on any real-world images?
- Will you ablate global max pooling vs global average pooling in teacher ResNet50 before distillation, following Islam et al.'s finding that max pooling outperforms average?
- What specific quantization method (PyTorch native, TensorFlow Lite, CoreML)—and will you measure calibration (ECE) separately on quantized student vs fp32 student?