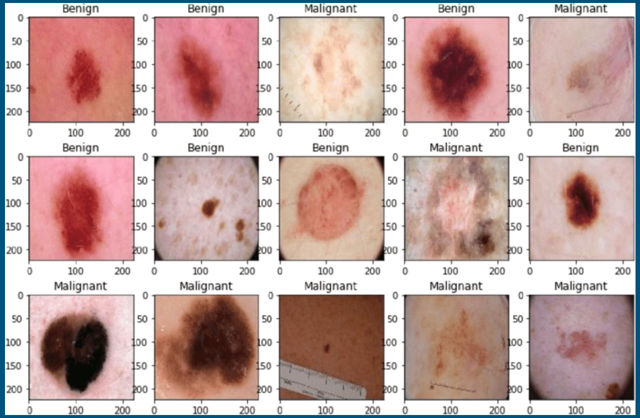
Optimizing On-Device Skin Cancer Detection

Theodore Villalva, Ryan Caudill, DJ Ravenell

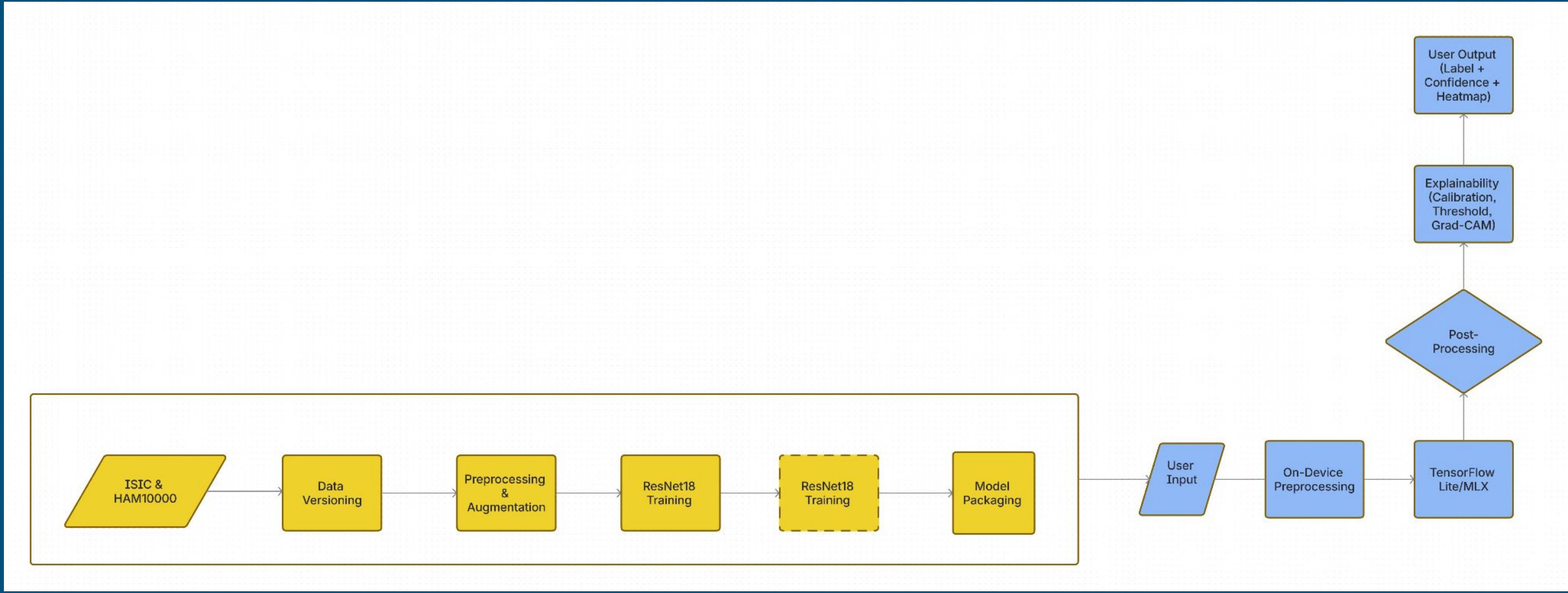
Problem

- Skin cancer is common; early detection saves lives.
  - 1/5
- Many people lack access to dermatologists or screening tools.

- Existing ML tools: research-only, cloud-based, or compute-heavy.
  - “Black box” interpretability
- How can we build a fast, lightweight, and interpretable skin lesion classification tool that performs reliably on small datasets and can run on-device for real-time use?



Solution

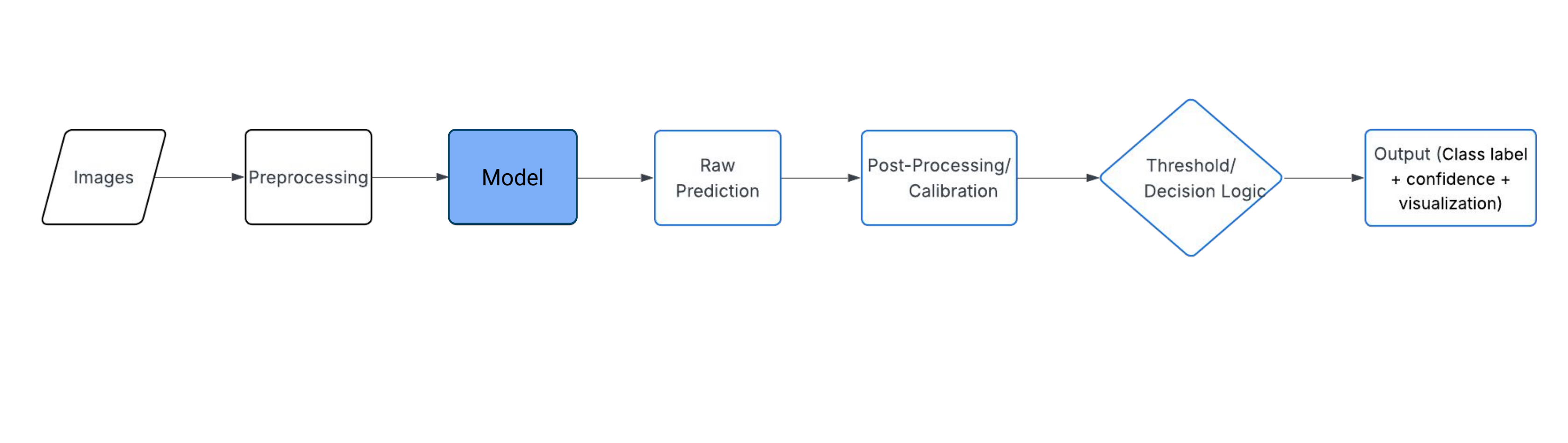


ISIC & HAM10000
Data Versioning
Preprocessing & Augmentation
ResNet18 Training
ResNet18 Training
Model Packaging
User Input
On-Device Preprocessing
TensorFlow Lite/MLX
Post-Processing
Explainability (Calibration, Threshold, Grad-CAM)
User Output (Label + Confidence + Heatmap)

Method

- **Model & Training**: Pre-trained CNN, fine-tuned via transfer learning on datasets ISIC and HAM10000 with preprocessing and augmentation.
- **Optimization**: Model quantized (8-bit) to reduce latency and memory usage for on-device usage.

Solution



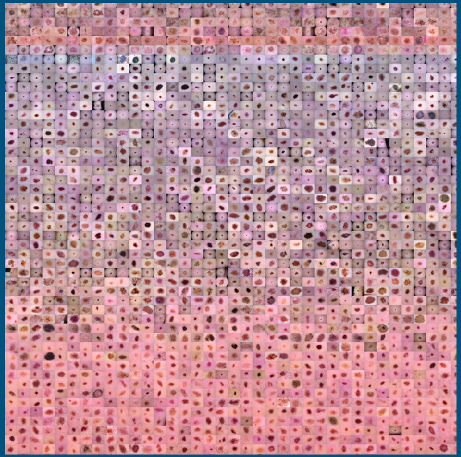
Images
Preprocessing
Model
Raw Prediction
Post-Processing/ Calibration
Threshold/ Decision Logic
Output (Class label + confidence + visualization)

Hypothesis

We hypothesize that an end-to-end pipeline for on-device skin cancer detection can leverage a pretrained CNN to achieve high diagnostic accuracy while reducing latency and memory footprint through 8-bit quantization. This optimization will enable a practical prototype for preliminary screening for edge devices such as M1 or M2 Macbooks.

Data

- Public datasets: ISIC (dataset of skin lesion images for dermatology research) and HAM10000 (10k dermatoscopic images)
- Includes labeled skin lesion images (melanoma, benign, etc.)
  Optional augmentation: lighting, rotation, zoom
- Supports training and evaluation of CNN models



Challenges

- Local compute may reduce model accuracy
- Skin lesions vary across skin tones → potential bias
- Lighting, angle, and image quality affect predictions
- Biopsy required for official diagnosis → model is preliminary only

Evaluation

- **ML Metrics**
  - Accuracy, Sensitivity, Specificity, F1 Score
- **Systems Metrics**
  - Full Precision vs. Quantized (8-bit) models
  - Inference latency & memory footprint
  - Energy usage
- **Explainability**
  - Grad-CAM heatmaps for interpretability