MoleCheck

Privacy-Preserving Skin-Lesion Analysis

Capstone Project Report

A mobile application that classifies a photographed skin lesion as lower- or higher-risk using a convolutional neural network, with all inference running on-device — no server, no upload, full offline operation.

Arel · RLcapstone.ai

1. Summary

MoleCheck classifies a photograph of a single skin lesion as lower-risk (benign) or higher-risk (malignant) using a convolutional neural network. Inference runs **entirely on-device**: the model is bundled inside the application, executes on the phone's own hardware, and functions in airplane mode. Because the image never leaves the device, the design provides a strong privacy guarantee for sensitive medical imagery.

Evaluated on a held-out test set of 1,722 images, the model achieves a **ROC-AUC of 0.914**, detecting **90% of malignant lesions (sensitivity)** while correctly clearing **75% of benign lesions (specificity)**.

> Not a medical device. MoleCheck is an educational demonstration of image classification. It cannot diagnose skin cancer or any other condition and must not inform a health decision. Anyone concerned about a mole or a skin change should consult a dermatologist.

2. Motivation

Skin cancer is among the most curable cancers when detected early and among the most dangerous when it is not. Access to dermatological screening is uneven, and most people cannot judge which lesions, if any, warrant professional review. The goal of this project was to build a complete pipeline end to end: train a classifier on real dermatological images, export it to run on a phone with no server, and deliver it inside a functioning mobile application. The project therefore spans machine learning, model deployment, and mobile development.

3. Dataset and labeling

Training used a public ISIC Archive collection of **11,720 dermatoscopic images** (a superset of the HAM10000 dataset). Each image's diagnosis field was mapped to a binary label — benign or malignant — with indeterminate cases excluded. The dataset is heavily imbalanced toward benign lesions, which informed both the training procedure and the choice of decision threshold.

Preventing data leakage

The same lesion is sometimes photographed more than once. Allowing multiple images of one lesion to fall on both sides of the split would let the model memorize specific lesions and report inflated performance. To prevent this, the data was partitioned so that **every image of a given lesion remains in a single split**.

Split	Lesions	Images	Malignant %
Training	6,121	8,123	18.8%
Validation	1,309	1,726	18.1%
Test	1,309	1,722	18.2%

4. Model and training

The classifier is **YOLO11s-cls**, the classification variant of Ultralytics' YOLO11 (5.4M parameters), fine-tuned for 40 epochs on an NVIDIA RTX 5060 Ti via **transfer learning** from a pretrained backbone. A pretrained model with a clean mobile-export path was chosen deliberately: a strong, reproducible baseline that deploys reliably is more valuable here than a bespoke architecture that is difficult to ship. Standard augmentation (flips, rotations, and color jitter) was applied so the model is robust to variation in lighting and orientation.

5. Results and threshold selection

Accuracy is a misleading metric on this task: because most lesions are benign, a trivial "always benign" classifier already attains roughly **82%** accuracy while detecting no cancers. The evaluation therefore emphasizes sensitivity (malignant lesions detected) and specificity (benign lesions correctly cleared).

Metric	Value	Interpretation
ROC-AUC	0.914	Overall separability of the two classes
Sensitivity	90.1%	Malignant lesions correctly flagged
Specificity	74.7%	Benign lesions correctly cleared

Prioritizing sensitivity

The default 0.5 decision threshold is inappropriate for a screening task, because the two error types are asymmetric: a false negative (a malignant lesion reported as benign) is far more costly than a false positive (a benign lesion referred for review). The operating threshold was set to **0.137**, the point at which the model detects 90% of malignant lesions while still clearing 75% of benign ones. The application is deliberately tuned to err toward recommending professional evaluation.

6. Deployment and verification

On-device inference required converting the PyTorch model to TensorFlow Lite. Because the direct export path is unsupported on Windows, the model was routed through ONNX and converted with onnx2tf. Conversion correctness was **verified rather than assumed**: the exported model was evaluated against the original PyTorch model and reproduced its predictions to within 0.001 probability, with identical decisions. The model was then integrated into a **Flutter** application (a single codebase targeting Android and iOS), gated behind a first-run disclaimer acknowledgement, and tested on an Android device.

7. v2 — closing the smartphone gap

The most consequential limitation of v1 is that it was trained on **dermatoscopic** images — captured through a clinical lens in contact with the skin — while the application receives an ordinary **smartphone photograph**. The 0.914 headline was therefore measured on a domain the application never encounters. v2 quantifies that mismatch and then corrects it.

The evaluation set is **PAD-UFES-20**: 2,298 smartphone photographs of skin lesions from a Brazilian screening program, each carrying a patient identifier. The data is split **by patient**, using v1's benign/malignant label space; actinic keratosis is excluded because v1's source data marked it indeterminate and dropped it, so scoring it would not be a fair test of the same model. This leaves 969 patients and 1,564 images.

Model	Evaluated on	ROC-AUC	Specificity at 90% sensitivity
v1 (dermoscopy-trained)	its own dermoscopy test	0.914	75%
v1 (dermoscopy-trained)	smartphone photographs	0.743	32%
v2 (fine-tuned on smartphone)	the same smartphone photographs	0.920	75%

On smartphone photographs the dermoscopy-trained model fell from ROC-AUC 0.914 to **0.743**, and at the 90%-sensitivity operating point its specificity collapsed to 32% — it flagged roughly two-thirds of benign lesions. Fine-tuning the same model on smartphone photographs restored performance to **ROC-AUC 0.920**, matching v1's original in-domain figure but now on the images the application actually receives, with specificity back to 75%. This smartphone-trained v2 model, with a 0.368 decision threshold, is what the application and browser demonstration now run.

> v2's 0.920 is the more meaningful number because it is measured on the domain the application operates in. Evaluating a model on the data it will really face — rather than the most favorable version of the task — is the principle applied throughout this portfolio.

8. Fairness across skin tones

A model with strong average performance can still fail unevenly, and in dermatology the recognized concern is degraded performance on darker skin. Because PAD-UFES-20 records a Fitzpatrick skin-type for many images, v2 can be audited for this directly. A **single global decision threshold** — the 90%-sensitivity operating point — is applied to every group, and the errors are examined for disparity.

Skin type (Fitzpatrick)	Images	Malignant	Sensitivity
I–II (lighter)	130	123	94.3%
III–IV (medium)	48	41	78.0%
V–VI (darker)	3	3	too few to assess

The model catches 94% of malignant lesions on lighter skin but only 78% on medium skin — a roughly 16-point sensitivity gap in the direction the literature warns of, with enough malignant cases in each group (123 and 41) for the difference to be a real signal. Two honest limitations bound the audit: the dataset contains almost no darker skin (three Fitzpatrick V images, no VI), so performance on the group most affected **cannot be evaluated at all**; and skin type was recorded mainly for biopsied, mostly malignant lesions, leaving too few labelled benign cases to measure per-group specificity

honestly.

> The uncomfortable but honest conclusion: v2 is measurably better at detecting cancer on lighter skin than on darker skin, and the available data cannot even test the darkest skin. Closing that gap requires more representative data, not a modelling adjustment.

9. Limitations

- Not a medical device — an educational project, not validated for clinical use.
- A 90% sensitivity target still implies roughly one in ten malignant lesions is missed, so a 'lower-risk' result is not reassurance.
- The shipped v1 model was trained on dermatoscopic images; v2 (above) measures the resulting smartphone-domain gap and closes it, but the two use different evaluation datasets and are not a like-for-like comparison beyond the shared label definition.
- Public dermatology datasets under-represent darker skin tones, so performance is not guaranteed to be uniform across skin types — a recognized equity concern in dermatology AI.

These limitations are precisely why the application always directs the user to consult a dermatologist.

10. Future work

- Extend training and evaluation to datasets that include darker skin (Fitzpatrick V–VI), which PAD-UFES-20 barely contains — the fairness audit above cannot assess the group most affected by the equity gap.
- Fuse the image with the clinical metadata PAD-UFES-20 provides (age, lesion site, whether it changed), the way a dermatologist reasons.
- Add an abstention option so low-quality images are declined rather than classified.
- Complete the iOS build and conduct informal user testing.