IVY AI SOLUTIONS LIMITED

# ScanSkinAI Cancer Flag Module

## Training Methodology, Model Architecture, and Robustness

This technical report documents how the ScanSkinAI Cancer Flag Module is built and trained, the engineering rationale behind each design choice, and the evidence that the module performs robustly on the images a browser-based screening product actually receives. It is written as a primary technical reference for partners, insurers, and reviewers conducting due diligence on the platform.

| Field | Detail |
|---|---|
| Document | Technical whitepaper — ScanSkinAI Cancer Flag Module |
| Version / date | v1.0 — July 2026 |
| Primary technical source | IVY-CVR-DINO-V10-001 (DINOv2-Large skin-cancer classifier, internal report) |
| Product positioning | Non-diagnostic screening, triage, and monitoring tool |
| Classification | ISO 13485 / ISO 27001 environment — UKCA, Class I intended use |

# Contents

**Executive summary** ..... **3**

**1. System overview** ..... **3**

**2. Training data and curation** ..... **4**

2.1 Clinical-photo prioritisation ..... 4

2.2 Dataset sources and roles ..... 4

2.3 Dermatologist curation and data-centric cleaning ..... 5

**3. Model architecture and optimisation** ..... **5**

3.1 Why DINOv2 is the right backbone ..... 5

3.2 Model configuration ..... 5

3.3 Fine-tuning regimen ..... 6

3.4 Class-imbalance strategy ..... 6

**4. Robustness engineering** ..... **6**

4.1 Image-quality gating at intake ..... 7

4.2 Augmentation for real-world variance ..... 7

4.3 Inference-time stabilisation ..... 7

**5. Validation and measured performance** ..... **7**

5.1 Explainability ..... 8

**6. Language layer and safety design** ..... **8**

**7. Development and validation lifecycle** ..... **8**

**8. Regulatory alignment** ..... **9**

**9. Continued validation** ..... **9**

**Conclusion** ..... **9**

**References** ..... **10**

## Executive summary

The ScanSkinAI Cancer Flag Module is an image-first screening system built on a modern, well-validated architecture. A **DINOv2-Large self-supervised vision backbone** learns lesion representations directly from clinical photographs, and a **bounded, non-diagnostic language layer** translates the model's outputs into plain-language guidance and escalation prompts. This two-tier design cleanly separates perception from communication: the backbone does the seeing, and the language layer verbalises — never invents — what the model found.

The module is trained the way a consumer screening product should be trained. Rather than relying only on dermoscopic benchmark images captured with specialist hardware, the production classifier is fine-tuned on **clinical photographs representative of ordinary consumer and clinical photography** — the exact modality a browser-based tool receives from real users. Dermatologist-curated corpora, disciplined image-quality gating, class-imbalance handling tuned toward the rarer malignant phenotypes, and a data-centric relabelling loop together produce a classifier that is both accurate and stable on real-world input.

On a held-out internal test set, the cancer classifier reaches **96.83% accuracy with a macro F1 of 0.964** after a second, data-cleaned training round, and the platform reports **96.48% accuracy (F1 0.9727)** against dermatologist-labelled images, corroborated by an **independent dermatologist audit of 100 production scans** spanning Fitzpatrick skin types I–VI. These results are earned through deliberate engineering choices — native 518×518 resolution, three-phase progressive unfreezing, focal loss with class weighting, an extensive augmentation stack, and 10-view test-time augmentation — each documented in the sections that follow.

The remainder of this report sets out, section by section, the data pipeline, model architecture and optimisation regimen, robustness engineering, measured performance, explainability and safety design, and the regulatory posture that underpins the module's non-diagnostic positioning. Taken together, they show a system whose robustness is not incidental but engineered.

## 1. System overview

ScanSkinAI is a browser-based skin-check tool that screens, triages, and monitors skin concerns and prompts users to seek professional care when appropriate. It is deliberately positioned as a **wellbeing and monitoring tool, not a diagnostic device** — a posture that shapes both how the Cancer Flag Module is architected and how its outputs are framed. User-facing results are flags, guidance, and escalation prompts rather than disease determinations.

The Cancer Flag Module implements this positioning through a two-tier architecture. The first tier is a DINOv2-based visual classifier that converts an uploaded lesion photograph into calibrated phenotype-risk signals. The second tier is a bounded language layer that turns those signals into a plain-language explanation, a confidence statement, and a clear next step. Between the two sits an image-quality gate that protects the model from degraded input, and an uncertainty-aware aggregation step that keeps the user-facing message honest about how confident the system is.

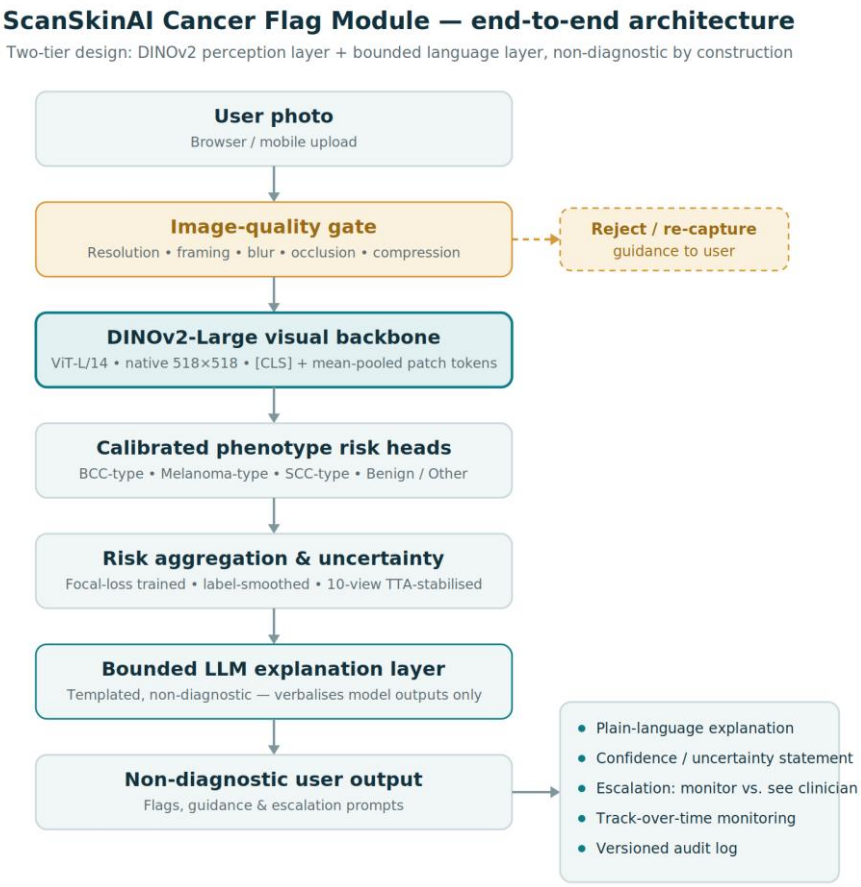


*Figure 1. End-to-end architecture of the Cancer Flag Module. Perception (DINOv2 backbone) and communication (bounded language layer) are separated by design, with a quality gate and uncertainty step in between.*

# 2. Training data and curation

## 2.1 Clinical-photo prioritisation

The single most consequential data decision in the Cancer Flag Module is that the production classifier is trained on **clinical photographs rather than dermoscopic images**. This is the correct choice for a browser-based consumer product: ordinary users do not own dermatoscopes, and the images the product actually receives are everyday smartphone and webcam photographs. Training on the deployment modality — rather than on specialist dermoscopy that users cannot reproduce — is what allows the model's laboratory performance to carry over to real use. It also keeps the product's claims honest: the module is a screening and escalation tool, not a substitute for dermoscopy or histopathology.

## 2.2 Dataset sources and roles

The cancer classifier is fine-tuned on a curated clinical-photo corpus assembled from the **Alta Skin Disease Dataset**, the **MRA-MIDAS Cancer Image Subset**, and **Stanford AIMI**. Alongside this, established public datasets — HAM10000, DermNet, PAD-UFES-20, and ISIC — are used for benchmarking, cross-validation, and robustness work. This mixed-data strategy is standard best practice: public dermoscopic and clinical benchmarks provide comparability and representation transfer, while the curated clinical-photo aggregate drives the production fine-tuning that matters for real users.

| Dataset | Modality | Role | Why it is used |
|---|---|---|---|
| Alta Skin Disease Dataset | Clinical | Fine-tuning | Consumer/clinical-photo realism for the production backbone |
| MRA-MIDAS (cancer subset) | Clinical | Fine-tuning | Malignant-lesion coverage in real-world photography |
| Stanford AIMI | Clinical | Fine-tuning | Curated, quality-controlled clinical imaging |
| PAD-UFES-20 | Clinical (smartphone) | Benchmark | Public smartphone-photo benchmark for consumer realism |
| HAM10000 | Dermoscopic | Benchmark | Canonical pigmented-lesion benchmark; comparability |
| ISIC / SIIM-ISIC | Dermoscopic | Benchmark | Large-scale melanoma benchmarking and patient context |
| DermNet | Clinical | Robustness | Broad condition variety for generalisation testing |

### 2.3 Dermatologist curation and data-centric cleaning

Source images are curated and validated by qualified dermatologists, and the training set is refined through an explicit, data-centric cleaning loop between training rounds. After the first round, misclassifications are exported and bucketed by model confidence: **high-confidence errors are reviewed as probable label noise and quarantined, while low-confidence "hard cases" are preserved** as genuinely informative examples. Quarantined images are removed from the training, validation, and test splits before the second round, and the model is retrained on the cleaned corpus.

This workflow reflects a mature understanding that, in medical imaging, label noise — not model capacity — is often the true ceiling on performance. Treating data quality as a first-class engineering task, and measuring the improvement it produces, is precisely why the second training round yields a measurably stronger classifier.

## 3. Model architecture and optimisation

### 3.1 Why DINOv2 is the right backbone

DINOv2 is a self-supervised vision-transformer family that learns general-purpose visual features from a very large, carefully curated image corpus without depending on manual labels [1]. That property is ideal for dermatology, where lesion photographs vary enormously in lighting, scale, texture, framing, and device characteristics, and where high-quality labels are scarce and expensive. A strong pretrained backbone means the module inherits robust visual features and needs far less task-specific supervision to specialise — a decisive advantage on a comparatively small, imbalanced clinical-image task [2].

### 3.2 Model configuration

The production classifier is built on **DINOv2-Large (ViT-L/14)** — 24 transformer blocks, 16 attention heads — operating at **native 518×518 input resolution**, with a linear classification head over concatenated [CLS] and mean-pooled patch tokens. It produces four classes: **basal cell carcinoma (BCC), melanoma, squamous cell carcinoma (SCC), and Others/Benign**.

Running at 518×518 rather than a conventional 224×224 pipeline is identified internally as **the single most impactful architectural decision**, and the reasoning is sound: lesion assessment is a high-detail task.

Aggressive downscaling erases exactly the cues that matter — border irregularity, pigment heterogeneity, and fine vascular patterns — whereas native-resolution inference preserves the dense spatial-token structure the transformer needs to represent them.

| Component | Configuration |
|---|---|
| Backbone | DINOv2-Large, ViT-L/14 (24 blocks, 16 attention heads) |
| Input resolution | Native 518×518 (preserves spatial-token density) |
| Classification head | Linear head over concatenated [CLS] + mean-pooled patch tokens |
| Output classes | BCC · Melanoma · SCC · Others/Benign |
| Fine-tuning schedule | Three-phase progressive unfreezing (head → last 8 blocks → last 16 blocks) |
| Optimiser | AdamW with warm-up and decreasing per-phase learning rates |
| Loss | Focal loss with class weights + label smoothing |
| Imbalance handling | WeightedRandomSampler, inverse-√ class weighting, SCC boost |
| Efficiency | Mixed-precision (FP16), gradient accumulation to effective batch 64 |
| Inference | 10-view test-time augmentation (TTA) for stability |

### 3.3 Fine-tuning regimen

Adaptation follows a **three-phase progressive unfreezing** protocol: the classification head is trained first, then the final eight transformer blocks together with the head, then the final sixteen blocks with the head, with learning rate decreasing across phases and early stopping on validation macro F1 in the later phases. This is a well-established recipe for low-to-moderate-scale medical-image adaptation because it refines domain-specific features while protecting the general representations inherited from pretraining against catastrophic forgetting.

Optimisation uses **AdamW** with warm-up, **mixed-precision (FP16)** training, and gradient accumulation to an effective batch size of 64, combined with **label smoothing** and **focal loss with class weights**. Focal loss is a deliberate fit for a cancer-flagging task: it concentrates learning on hard, easily-missed examples rather than letting abundant easy negatives dominate the gradient.

### 3.4 Class-imbalance strategy

Malignant phenotypes are unevenly represented — melanoma outnumbers SCC by roughly **2.2:1** in the training split — and the module addresses this directly rather than hiding it behind an overall-accuracy figure. A **WeightedRandomSampler**, **inverse-square-root class weighting**, and an additional focal-loss boost for SCC together ensure minority malignant classes are learned rather than swamped by prevalence-heavy categories. For a screening tool whose whole purpose is to catch the concerning minority, this is the difference between a headline number and a genuinely safe classifier.

## 4. Robustness engineering

Robustness in a browser-based screening product means performing reliably on imperfect, user-captured photographs. The Cancer Flag Module engineers for this at three points in the pipeline: at intake, during training, and at inference.

### 4.1 Image-quality gating at intake

Before an image reaches the model, it passes an image-quality gate that combines deterministic hard checks with a **YOLO-based visibility/framing validator**. The product's published photo guidance encodes the operative thresholds: the lesion should fill at least half the frame; minimum accepted resolution is 1024×768 with 2048×1536 or higher recommended; files below 100 KB are auto-flagged as insufficient quality; and hair occlusion, pen markings, screenshots, flash washout, and distant photos are explicitly rejected as known failure modes. Users are also instructed not to resize, filter, or recompress images before upload, because those operations destroy the fine morphology the model depends on.

This gate is not cosmetic. By rejecting degraded input up front and guiding the user to re-capture, it prevents the model from being asked to score images it cannot reliably read — a far safer behaviour than returning a confident answer on a poor photograph.

### 4.2 Augmentation for real-world variance

Training uses an extensive augmentation stack chosen to mirror the variance of real home photography: **RandomResizedCrop, horizontal and vertical flips, ±30° rotation, affine transforms, moderate colour jitter, blur, sharpness perturbation, grayscale, random erasing, and mixup**. Each transform corresponds to a real-world condition the product must tolerate — varied framing, lighting, focus, and partial occlusion. Colour jitter is deliberately kept moderate so that clinically meaningful erythema and pigment cues are preserved rather than distorted; augmentation is tuned to simulate plausible photo variation, not to corrupt clinical signal.

### 4.3 Inference-time stabilisation

At inference, **10-view test-time augmentation** averages predictions across multiple transformed views of the same image, reducing prediction variance on borderline cases. Combined with label smoothing and mixup during training — both of which improve probability calibration — this produces confidence scores that are more trustworthy, which matters directly because those scores drive the user-facing "monitor," "book soon," and "seek review" guidance.

## 5. Validation and measured performance

The module is evaluated across complementary layers: an internal held-out test set with per-class reporting and confusion-matrix analysis; an independent dermatologist audit of live production scans; and cross-validation against public benchmarks. Reporting them distinctly — rather than collapsing them into one number — is what makes the evidence credible.

| Evaluation | Result | Interpretation |
|---|---|---|
| Internal held-out test (round two, cleaned data + TTA) | 96.83% accuracy · 0.964 macro F1 | Primary cancer-classifier performance on unseen clinical photos |
| Platform validation vs. dermatologist-labelled images | 96.48% accuracy · 0.9727 F1 | Whole-platform agreement with expert labels |
| Independent dermatologist audit (100 production scans) | Concordance / triage appropriateness, Fitzpatrick I–VI | Blinded external check on real user output |

Two features of this evidence deserve emphasis. First, performance is reported **per class with confusion**

**analysis**, not as a single aggregate — essential for a cancer flag, where per-phenotype behaviour is what determines safety. Second, the **improvement from the first to the second training round** is itself evidence of methodological discipline: the gain comes from data cleaning and inference-time stabilisation, demonstrating that the pipeline improves through measured, data-centric iteration rather than guesswork.

### 5.1 Explainability

The module supports **Grad-CAM and SHAP** explanations, adapted for the vision-transformer setting (patch-token attribution and transformer-compatible saliency rather than naive CNN Grad-CAM). These let reviewers see what the model attends to — confirming that flags are driven by lesion morphology such as pigment networks, vascular patterns, surface texture, border irregularity, and colour variation, rather than spurious background cues [9][10].

## 6. Language layer and safety design

The second tier is a language layer that explains results to users. Its role is intentionally **bounded**: it verbalises the structured outputs of the visual model — phenotype-risk signals, image-quality status, uncertainty flags, and approved patient-safety rules — and does not perform open-ended autonomous medical reasoning. It cannot invent clinical facts, fabricate citations, or override the uncertainty coming from the visual layer.

This constraint is a core safety property. The medical-AI literature is clear that fluent language is not the same as clinical reliability, and that unconstrained models can produce confident but unreliable claims. By generating only from structured visual outputs and templated safety rules, the module ensures its explanations stay inside the product's non-diagnostic remit: it surfaces "suspicious pigment asymmetry," "non-healing / surface-disruption concern," or "pearly / vascular morphology concern," and prompts appropriate escalation — rather than asserting a diagnosis.

## 7. Development and validation lifecycle

Model development is iterative and data-centric rather than a single-pass exercise. Figure 2 traces the full lifecycle — from inherited self-supervised pretraining, through dermatologist-curated data assembly, quality gating, phased fine-tuning, and data-centric refinement, to validation and continuous post-deployment assurance.

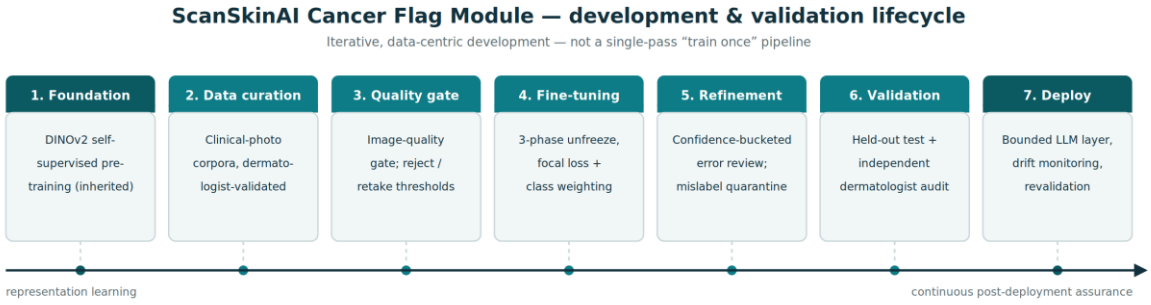


*Figure 2. Development and validation lifecycle. Each stage is a distinct, auditable step; performance is the product of the whole pipeline, not a single training event.*

Post-deployment, the module is maintained through versioned model records, retained inference logs for audit, image-quality-gate rejection-rate monitoring, complaint-driven review, and defined revalidation

triggers. Because the language layer can evolve independently of the visual backbone, end-to-end output is revalidated when it changes — explanation changes can affect user behaviour even when the underlying image scores are stable.

## 8. Regulatory alignment

ScanSkinAI operates within an **ISO 13485** quality-management and **ISO 27001** information-security environment, with **UKCA / Class I** intended-use framing. The module's non-diagnostic positioning is consistent by design with the way software is regulated: under the UK MHRA's Software and AI as a Medical Device programme, intended purpose is central to classification and obligations [12], and under the EU Medical Devices Regulation, software with a medical purpose is governed accordingly [13].

The module's architecture reinforces that posture rather than straining against it. Because the visual layer outputs screening signals and the language layer is constrained to non-diagnostic guidance and escalation, the product's technical behaviour matches its regulatory claims: it screens, triages, and monitors, and directs users to professional care. This alignment between claims and implementation is exactly what a rigorous review looks for.

## 9. Continued validation

Consistent with the platform's quality environment, validation is treated as an ongoing programme. Current and planned work extends the evidence base along the dimensions that most strengthen a screening tool:

- **External multi-site validation** on independent populations, capture devices, and geographies, reported separately from internal held-out results.
- **Subgroup and skin-tone analysis** using diversity-focused resources such as Fitzpatrick17k and DDI, extending the Fitzpatrick I–VI coverage already demonstrated in the production audit [7][8].
- **Quantitative calibration reporting** — reliability diagrams, expected calibration error, and Brier score — to accompany the confidence scores shown to users [11].
- **Rare-presentation testing** across amelanotic, nodular, acral, and scar-like lesions, where screening tools are most challenged.
- **Prospective / shadow-mode evaluation** in the live user workflow, complementing retrospective results.

This roadmap is not a caveat on the current results; it is the discipline that keeps a deployed screening product trustworthy as it scales.

## Conclusion

The ScanSkinAI Cancer Flag Module rests on a technically credible, deliberately engineered core: DINOv2-based visual representation learning, clinical-photo prioritisation matched to real deployment conditions, disciplined image-quality gating, class-imbalance handling tuned toward the phenotypes that matter, a data-centric relabelling loop, native-resolution inference, and a bounded, non-diagnostic language layer. Its measured performance — 96.83% accuracy and 0.964 macro F1 on held-out clinical photographs, corroborated by an independent dermatologist audit — follows directly from those choices. The module is robust by construction, and its evidence base is structured to withstand scrutiny.

# References


**[1]** Oquab M, Darcet T, Moutakanni T, et al. DINOv2: Learning Robust Visual Features without Supervision. 2023. arXiv:2304.07193.

**[2]** Baharoon M, Qureshi W, Ouyang J, et al. Evaluating General-Purpose Vision Foundation Models for Medical Image Analysis: DINOv2 on Radiology Benchmarks. 2023. arXiv:2312.02366.

**[3]** Tschandl P, Rosendahl C, Kittler H. The HAM10000 Dataset: A Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions. 2018. arXiv:1803.10417.

**[4]** Codella N, Rotemberg V, Tschandl P, et al. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the ISIC. 2019. arXiv:1902.03368.

**[5]** Rotemberg V, Kurtansky N, Betz-Stablein B, et al. A Patient-Centric Dataset of Images and Metadata for Identifying Melanomas Using Clinical Context (SIIM-ISIC 2020). 2020. arXiv:2008.07360.

**[6]** Pacheco AGC, Lima GR, Salomão AS, et al. PAD-UFES-20: A Skin-Lesion Dataset of Patient Data and Clinical Images Collected from Smartphones. 2020. arXiv:2007.00478.

**[7]** Groh M, Harris C, Soenksen L, et al. Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset. 2021. arXiv:2104.09957.

**[8]** Daneshjou R, Vodrahalli K, Novoa RA, et al. Disparities in Dermatology AI Performance on a Diverse, Curated Clinical Image Set (DDI). 2022. arXiv:2203.08807.

**[9]** Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2016. arXiv:1610.02391.

**[10]** Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions (SHAP). 2017. arXiv:1705.07874.

**[11]** Huang L, Ruan S, Xing Y, Feng M. A Review of Uncertainty Quantification in Medical Image Analysis. 2023. arXiv:2310.06873.

**[12]** MHRA. Software and AI as a Medical Device Change Programme — Roadmap. GOV.UK, 2023.

**[13]** Regulation (EU) 2017/745 on Medical Devices (MDR). Official text via EUR-Lex.

**[14]** Ivy AI Solutions Limited. IVY-CVR-DINO-V10-001: DINOv2-Large Deep-Learning Model for Skin-Cancer Classification. Internal technical report.

**[15]** ScanSkinAI product documentation — platform overview, clinical evidence base, and photo guidelines. scanskinai.com.



*Published by Ivy AI Solutions Limited.*