Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An offline smartphone skin-triage model that failed external validation: a negative-result report

Muhammad Uzair

muzair0908@gmail.com

Independent Researcher https://orcid.org/0009-0006-7837-3087

Research Article

Keywords: dermatology, artificial intelligence, external validation, skin of color, Fitzpatrick skin type, abstention, negative result, mobile health

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11595470/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Smartphone tools that place skin lesions into triage bands are often proposed for areas with too few dermatologists. Yet published studies tend to focus on internal performance, while external validation remains uncommon, particularly for pigmented skin. I evaluated a frozen, fully offline Android triage model (dermedge-1.8.0; MobileNetV2, INT8-quantized TensorFlow Lite, an eight-class classifier trained on HAM10000, PAD-UFES-20, and Fitzpatrick17k) using a scoring-only protocol on three external datasets not used for training: DDI, DDI-2, and MRA-MIDAS. The app converts the model's malignancy score into triage bands, including REFER. An uncertainty gate returns "cannot assess" rather than displaying a band when output is uncertain. In total, 3,266 images were scored, of which 3,034 had usable labels. On the model's own held-out test split, lesion-only AUC was 0.948. External in-scope AUCs were 0.683 [0.624, 0.740] for DDI, 0.719 [0.632, 0.805] for DDI-2, and 0.692 [0.657, 0.727] for MRA-MIDAS. Since the intervals overlap, the results indicate similar performance, around 0.68 to 0.72, rather than a ranking across datasets. For in-scope DDI images labelled Fitzpatrick V–VI (81 images, 16 malignant), AUC was 0.588 [0.404, 0.753], indistinguishable from chance. The app returned "cannot assess" for 72–80% of external images and displayed REFER for 7–18%. Among 1,057 malignant images, 808 received "cannot assess" and 41 appeared in the reassuring LOW band. The model failed external validation by a wide margin and is not ready for a pilot. I report these results in full because the internal–external gap, dark-skin performance, and abstention burden are central to any responsible deployment decision. No clinical pilot should proceed until a frozen model meets a pre-registered external-performance bar.

1. Introduction

Rural Pakistan, like many parts of sub-Saharan Africa and South Asia, has far fewer dermatologists than the population needs. A smartphone tool that places skin lesions into triage bands, from referral to reassurance, could help frontline health workers. Its value depends on whether performance holds on images outside the training data and across skin tones that are underrepresented in those data.

I developed DermEdge AI, a fully offline Android app that photographs a skin lesion and assigns a triage band. The app is intended only as a research instrument. In any future pilot, the clinician would assess the lesion first, then take a photograph, and record the app's output without acting on it. Here, I report the results from scoring the frozen model against three external benchmarks. They were negative, and they explain why no pilot is planned.

Negative findings such as these appear infrequently in dermatology AI, leaving the literature weighted toward successful internal validations. This paper documents one failure in full, including the difference between internal and external performance, results on dark skin, and the burden created by abstention. The methods are described in enough detail to help others avoid the same path.

2. Related Work

Freeman et al. reviewed diagnostic-accuracy studies of algorithm-based smartphone apps for skin cancer. They found few studies and a high risk of bias, and concluded that these apps could not yet be relied on to detect skin cancer. They also warned that performance might be worse when intended users use the apps than in research settings (1). Daneshjou et al. reviewed dermatology AI studies and found that the datasets behind them were rarely public and rarely described the skin tones represented (2). The group later assembled the Diverse Dermatology Images (DDI) set and showed that state-of-the-art dermatology models performed substantially worse on this dataset than on their original test data, with the largest declines for dark skin and uncommon diseases (3). Groh et al. introduced Fitzpatrick17k, a clinical image dataset with Fitzpatrick labels, and found that models trained on some skin types were less accurate on others (4). This paper adds a fully reported failure of an offline model intended for deployment, evaluated on three external datasets, with abstention reported alongside AUC.

3. Methods

3.1 Model and app

The evaluated model, dermedge-1.8.0, uses a MobileNetV2 network (5), quantized to INT8 TensorFlow Lite, and was trained as an eight-class skin-lesion classifier (akiec, bcc, bkl, df, mel, normal, nv, vasc). Training used three public datasets: HAM10000 (6), containing 10,015 dermatoscopic images from clinics in Austria and Australia and roughly two thirds of the lesion images; PAD-UFES-20 (7), with 2,298 smartphone photographs from Brazil; and a 2,588-image subset of Fitzpatrick17k (4), containing clinical photographs with Fitzpatrick labels. The "normal" class consists of 2,500 healthy-skin crops extracted from PAD-UFES-20 and Fitzpatrick17k photographs. The malignancy score is the summed probability for the three malignant classes (akiec, bcc, mel), excluding "normal" (no lesion). The shipped thresholds define LOW as scores below 0.234375, INDETERMINATE as scores from 0.234375 to below 0.363281, and REFER as scores at or above 0.363281. On the training corpus's validation split, the rule-out boundary was set for sensitivity at or above 0.95, and the refer boundary for precision at or above 0.70. These boundaries are provisional. They have not been validated on deployment data or approved by a clinician.

The app runs fully on-device and does not depend on a network connection. I froze the dermedge-1.8.0 weights before external scoring, and no image from external validation was used for training. Every version from dermedge-1.3.0 onward that underwent an external test, including this version, was evaluated on DDI during development. DDI therefore informed development decisions and is partly a development set. DDI-2 and MRA-MIDAS were first obtained after the model was frozen.

3.2 External evaluation protocol

I scored the model, without retraining, on three public benchmarks. None contributed training data:

DDI (Diverse Dermatology Images) (3): biopsy-proven images with Fitzpatrick skin-type labels, including the dark-skin subset (Fitzpatrick V–VI).

DDI-2 (8): a later release of DDI containing diverse skin-condition images from self-identified Asian patients.

MRA-MIDAS (9), the Melanoma Research Alliance Multimodal Image Dataset for AI-based Skin Cancer, a multimodal benchmarking dataset. I scored only its clinical photographs. The 1,049 dermoscopic images were excluded because the app uses a phone camera, not a dermatoscope. Most lesions have two clinical photographs (929 of 962), so confidence intervals for this dataset resample whole patients.

Across all three datasets, 3,266 images were scored, and 3,034 had usable labels. These figures count images, not lesions.

Before displaying a band, the app applies a fixed anti-fabrication gate. It shows "Cannot assess this image" if the top class probability is below 0.55 or the normalised entropy of the output distribution exceeds 0.80. When the gate is triggered, the band is still calculated and recorded but hidden from the health worker. The external evaluation used the same gate, so the abstention rates reported below approximate what a health worker would see (see Limitations).

3.3 Metrics

The primary metric was AUC, reported with 95% percentile bootstrap confidence intervals (2,000 resamples, seed 42). I calculated AUC on in-scope images, defined as diagnoses represented by one of the model's classes, using the same label table across all three datasets. Actinic keratosis was excluded because labels differ across datasets. If a dataset's label conflicted with the table, I excluded the image rather than relabelling it. AUC was calculated from the continuous score for every in-scope image, including images for which the gate would hide the band. I also report AUC across all labelled images, including diagnoses outside the model's classes. Confidence intervals for DDI-2 and MRA-MIDAS resampled whole patients. Because DDI has no patient key, its interval resampled images and is somewhat too narrow.

To assess clinical burden and safety, I also report the cannot-assess (abstention) rate, the share of images shown with REFER, and the displayed bands for malignant images. A tool that abstains on most images will not reduce referral load, regardless of its AUC on the remaining images. Dark-skin performance was evaluated on in-scope DDI images labelled Fitzpatrick V–VI, using the same label table as the other analyses. Under DDI's own labels, which classify one additional image as malignant (an atypical spindle-cell nevus of Reed, which the table classifies as a nevus), the same subset yields 0.577 (82 images, 17 malignant).

4. Results

On the model's held-out test split, lesion-only AUC was 0.948. The split was drawn from the model's training corpora using a stratified 80/10/10 division into training, validation, and test sets (StratifiedGroupKFold, seed 42). Grouping was by lesion for HAM10000, by patient for PAD-UFES-20, and by image for Fitzpatrick17k, which has no patient key, so each group appeared in only one subset. The

test split contained 1,766 images. After removing its 260 healthy-skin crops, 1,506 lesion images remained, including 560 malignant images. AUC was measured on the floating-point model before INT8 conversion. Across all 1,766 test images, conversion changed AUC from 0.9549 to 0.9535. I did not compute a confidence interval for the internal figure. Earlier project documents reported an internal AUC of 0.961, but it came from a later comparison run with different healthy-skin mining, which redrew the split. That test split included lesions the model had been trained or tuned on, so it was not held out and is not reported here. The internal figure appears only alongside external results, which are the main finding.

Table 1

In-scope AUC of dermedge-1.8.0 across the three external datasets, alongside AUC for all labelled images. The 95% percentile bootstrap intervals are based on 2,000 resamples. DDI-2 and MRA-MIDAS resample whole patients, whereas DDI resamples images because no patient key is available. The external confidence intervals overlap substantially. Together, they indicate performance around 0.68 to 0.72 rather than a meaningful ranking among datasets. When diagnoses outside the model's classes are included, all AUCs fall to 0.63 to 0.67. Since DDI informed model development, DDI-2 and MRA-MIDAS provide cleaner external tests. Their scores were 0.719 and 0.692, against DDI's 0.683. This pattern does not suggest that selecting models on DDI inflated the DDI result.

Dataset	**In-scope images (malignant)**	**In-scope AUC [95% CI]**	**All labelled images (malignant)**	**All-diagnosis AUC [95% CI]**
DDI	396 (123)	0.683 [0.624, 0.740]	651 (170)	0.650 [0.598, 0.698]
DDI-2	286 (43)	0.719 [0.632, 0.805]	645 (70)	0.632 [0.555, 0.708]
MRA-MIDAS	1,410 (807)	0.692 [0.657, 0.727]	1,738 (817)	0.665 [0.630, 0.696]

Table 2

In-scope AUC on DDI by Fitzpatrick group. Image bootstrap, 2,000 resamples. The intervals are somewhat too narrow because DDI has no patient key. For in-scope DDI images labelled Fitzpatrick V–VI (81 images, 16 malignant), AUC was 0.588 [0.404, 0.753]. This interval was the only one of the three Fitzpatrick groups to include chance, and it was the widest, reflecting the 16 cancers in this subset. Across all 206 DDI V–VI images, including diagnoses not represented among the model's classes (47 malignant), AUC was 0.594 [0.498, 0.686]. DDI-2 and MRA-MIDAS contain no Fitzpatrick V–VI cancers, leaving DDI as the only source for this measurement.

DDI, in scope	Images (malignant)	AUC [95% CI]
Fitzpatrick I–II	133 (42)	0.617 [0.511, 0.714]
Fitzpatrick III–IV	182 (65)	0.745 [0.665, 0.820]
Fitzpatrick V–VI	81 (16)	0.588 [0.404, 0.753]

Table 3

The bands shown by the app for each scored image, with the gate applied first. The lower block counts malignant images with usable labels under the same label table, including diagnoses outside the model's classes.

All scored images	DDI	DDI-2	MRA-MIDAS
	656	**665**	**1,945**
"Cannot assess"	474 (72.3%)	523 (78.6%)	1,546 (79.5%)
Displayed REFER	116 (17.7%)	49 (7.4%)	246 (12.6%)
Displayed INDETERMINATE	24 (3.7%)	30 (4.5%)	66 (3.4%)
Displayed LOW	42 (6.4%)	63 (9.5%)	87 (4.5%)
Malignant images	170	70	817
"Cannot assess"	117	51	640
Displayed REFER	41	8	138
Displayed INDETERMINATE	4	5	12
Displayed LOW	**8**	**6**	**27**

Abstention was the main operational constraint. The app returned "cannot assess" for 72.3% of DDI images (474 of 656), 78.6% of DDI-2 images (523 of 665), and 79.5% of MRA-MIDAS images (1,546 of 1,945). Across the datasets, this amounted to 72–80% of all scored images. The app displayed REFER for 17.7% (116 of 656), 7.4% (49 of 665), and 12.6% (246 of 1,945), or 7–18%. If the gate were ignored and only the bands were considered, 64–79% of images would fall in REFER. That describes the underlying scores, not the results shown to users. A tool unable to assess roughly three images in four,

while referring as many as about one in five, does not perform useful triage. Instead, it adds a step. Of the 1,057 malignant images, 808 (76%) received "cannot assess", while 41 were shown as LOW, the reassuring result: 8 on DDI, 6 on DDI-2, and 27 on MRA-MIDAS.

5. Discussion

Three findings are especially relevant for teams developing tools of this kind.

The gap between internal and external performance is the main finding. An internal AUC of 0.948 alongside external AUCs of 0.68 to 0.72 is not a minor overestimate. It suggests that the distribution used in development differs from the one the model encounters in external evaluation. Reports that foreground the internal result risk giving a misleading picture, so this paper places the external results first. About two thirds of the training lesion images are dermatoscopic, while external images are clinical photographs, making a shift in imaging modality an obvious possible explanation. If modality were the main cause, dermoscopic images of the same lesions should have scored higher. On 746 MRA-MIDAS lesions (456 malignant), they did not. AUC was 0.679, compared with 0.711 for one clinical photograph, a difference of − 0.032 [− 0.078, + 0.013] (patient bootstrap, 5,000 resamples; this paired comparison uses MRA-MIDAS's own labels, which count actinic keratosis as malignant).

For dark skin, the model's performance was indistinguishable from chance: AUC was 0.588, with a 95% confidence interval of 0.404 to 0.753, based on 16 cancers. These data do not establish that the model performs worse on dark skin than on light skin. DDI images in Fitzpatrick I–II had an AUC of 0.617 [0.511, 0.714], and the dark-skin interval also includes the overall DDI result. The available dark-skin cancer images are too few to show whether the model ranks those cancers better than chance. For a tool intended for settings where darker skin is common, this uncertainty is itself disqualifying, regardless of the overall AUC.

Abstention is a clinical outcome, not a minor implementation detail. Reporting AUC only for assessed images, without also stating the 72–80% cannot-assess rate, would make the model appear more useful than it is. The abstention rate determines whether the tool reduces or adds to frontline workload. The safety results point in the same direction: three in four malignant images received no verdict, and 41 were displayed as LOW, precisely the outcome triage is meant to prevent.

Retraining on the same datasets did not close the gap. Around dermedge-1.8.0, I made six training-side changes, one at a time. These included changes to healthy-skin crop mining, which were dropped after performing worse internally and before any external test; selecting checkpoints on the worst-performing corpus rather than pooled validation AUC; lowering the brightness floor and then raising the saturation ceiling for healthy-skin mining, with the latter change producing dermedge-1.8.0; and adding a ninth class for non-neoplastic disease in two separate runs. None of the five changes scored on DDI improved external performance. Adding the ninth class first reduced performance and then, after a scoring fix, recovered without exceeding dermedge-1.8.0. Selecting the checkpoint on the worst-performing corpus

raised that corpus's validation AUC from 0.78 to 0.91 but had no effect on DDI. Robustness across the datasets available for development did not translate into robustness on an unseen dataset.

The clinician-first protocol also needs a clear interpretation. The clinician's assessment is recorded before the app's output is revealed, and the output is not acted on. This design preserves an independent clinical judgment, but it does not show whether AI assistance improves or harms decisions. Answering that requires a separate study arm in which the output informs the decision. This distinction matters because the current protocol does not test that question.

On the basis of these results, I will not run a clinical pilot until a frozen model meets a pre-registered external-data threshold, including performance stratified by dark skin and a bounded abstention rate. The next planned step is a scoring-only evaluation on the PASSION benchmark (10). The access request form was submitted on 2026-09-09 and is awaiting human review, with no response as of this writing. PASSION will be used for scoring only, not training.

For small teams, the lesson is procedural: freeze the model, evaluate it on predefined test splits never used during training, stratify results by skin tone using each dataset's own labels, and report abstention beside AUC. The internal test split must also be frozen and recorded. The project's first internal figure, 0.961, came from a split re-derived in a later run, which meant the result was not based on a held-out test set. Freezing the model alone would not have prevented this problem. Without those steps, the paper would have reported only the internal figure.

6. Limitations

This retrospective study evaluates image benchmarks rather than a prospective clinical setting. Real-world capture conditions, such as lighting, camera quality, and how health workers select lesions, might differ further.

I evaluated one frozen model version. The six training-side changes described above used the same three datasets, so they do not show what training on substantially different data, especially images of cancers on dark skin, would achieve. Retraining on these benchmarks was outside the study's scope because the evaluation was designed for scoring only.

Fitzpatrick labels are an imperfect proxy for skin tone.

The image-resizing method in the Android app differs from the one used for external scoring. With an earlier model version (dermedge-1.3.0), a resize approximating the Android pipeline increased DDI abstention from 61.6% to 62.4%. The reported rates may therefore slightly understate what an Android phone would display. This comparison has not been repeated with dermedge-1.8.0.

Abstention also varies across training runs. The gate thresholds absolute confidence, and confidence values can shift between runs even when rankings barely change. Two identical trainings with this configuration produced DDI abstention rates of 63.1% and 72.3%. The shipped model is the second run.

A single abstention rate in this paper should therefore be interpreted as uncertain by about nine points in either direction.

The internal result has three additional limitations. It was measured on the floating-point model rather than the INT8 model, and no confidence interval was calculated. Fitzpatrick17k also lacks a patient key, so the split was made by image. This removes exact copies but does not rule out near-duplicate photographs of the same lesion. If such images fell into both training and test sets, the internal AUC would be inflated. These limitations do not affect the external results.

7. Conclusion

dermedge-1.8.0 did not pass external validation. Its overall in-scope AUC was 0.68 to 0.72, its AUC on dark skin was 0.588 and indistinguishable from chance, and its abstention rate was 72–80%. The model is not ready for a clinical pilot, and I will not pilot it. I am reporting these results so the evidence behind this decision is available to the next team.

Declarations

Data and Code Availability

On reasonable request, the author will provide the evaluation harness and frozen-model definition. DDI, DDI-2, and MRA-MIDAS are available from their custodians under their respective licenses. No new patient data were collected for this study.

Ethics Statement

This study used only existing, de-identified image datasets, obtained from their custodians under their research-use terms. No new patient data were collected, and no human subjects were enrolled or contacted.

Use of AI Tools

Claude (Anthropic), an AI assistant, was used in this work. The original idea was the author's, and the architecture, design, and backbone of the app were planned and architected by the author. Claude assisted with bug fixing and auditing the code. The author ran the scoring himself. The results, interpretation, and decisions in this paper are the author's own.

Conflicts of Interest and Funding

The author declares no conflicts of interest. This work received no external funding.

References

1. Freeman, K., Dinnes, J., Chuchu, N., Takwoingi, Y., Bayliss, S. E., Matin, R. N., Jain, A., Walter, F. M., Williams, H. C., & Deeks, J. J. (2020). Algorithm based smartphone apps to assess risk of skin cancer in adults: Systematic review of diagnostic accuracy studies. *BMJ*, 368, m127. https://doi.org/10.1136/bmj.m127
2. Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V., & Zou, J. (2021). Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review. *JAMA Dermatology*, 157(11), 1362–1369. https://doi.org/10.1001/jamadermatol.2021.3129
3. Daneshjou, R., Vodrahalli, K., Novoa, R. A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S. M., Bailey, E. E., Gevaert, O., et al. (2022). Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, 8(32), eabq6147. https://doi.org/10.1126/sciadv.abq6147
4. Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., & Badri, O. (2021). Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1820–1828). IEEE. https://doi.org/10.1109/CVPRW53098.2021.00201
5. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4510–4520). IEEE. https://doi.org/10.1109/CVPR.2018.00474
6. Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5, 180161. https://doi.org/10.1038/sdata.2018.161
7. Pacheco, A. G. C., Lima, G. R., Salomão, A. S., Krohling, B., Biral, I. P., de Angelo, G. G., Alves, F. C. R., Jr., Esgario, J. G. M., Simora, A. C., Castro, P. B. C., Rodrigues, F. B., Frasson, P. H. L., Krohling, R. A., Knidel, H., Santos, M. C. S., do Espírito Santo, R. B., Macedo, T. L. S. G., Canuto, T. R. P., & de Barros, L. F. S. (2020). PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 32, 106221. https://doi.org/10.1016/j.dib.2020.106221
8. Chang, C. T., Pathmarajah, P., Allerup, J., Jafry, S., Yekrang, K., Mitchell, D. C., See, N. A., Perrone, L. A., Fong, B., Cisneros, M. D., Daneshjou, R., Ko, J., & Chiou, A. S. (2025). DDI-2: A diverse skin condition image dataset representing self-identified Asian patients. *Journal of Investigative Dermatology*, 145(5), 1205–1208.e4. https://doi.org/10.1016/j.jid.2024.09.018
9. Chiou, A. S., Omiye, J. A., Gui, H., et al. (2025). Multimodal image dataset for AI-based skin cancer (MIDAS) benchmarking. *NEJM AI*, 2(6), AIdbp2400732. https://doi.org/10.1056/AIdbp2400732 (Preprint: medRxiv, https://doi.org/10.1101/2024.06.27.24309562)
10. Gottfrois, P., Gröger, F., Andriambololoniaina, F. H., Amruthalingam, L., Gonzalez-Jimenez, A., Hsu, C., Kessy, A., Lionetti, S., Mavura, D., Ng'ambi, W., Ngongonda, D. F., Pouly, M., Rakotoarisaona, M. F., Rapelanoro Rabenja, F., Traoré, I., & Navarini, A. A. (2024). PASSION for dermatology: Bridging the diversity gap with pigmented skin images from sub-Saharan Africa. In *Medical Image Computing and*

Computer Assisted Intervention – MICCAI 2024 (Lecture Notes in Computer Science, Vol. 15003, pp. 703–712). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72384-1_66