

Towards Reliable and Fair Deep Learning for Skin Cancer Screening: An Explainable, Data-Aware Framework

Gavkhare Abhijeet Santosh¹, Prof. P. V. Gaikwad², Dr. H. B. Jadhav³

Student, Department of Computer Engineering¹

Guide, Department of Computer Engineering²

Co-Guide, Department of Computer Engineering³

Adsul Technical Campus, Chas, Ahilyanagar, India

Abstract: *Deep learning has become a useful research direction for computer aided analysis of skin lesions, but strong results on a single dataset do not necessarily translate into reliable clinical performance. This paper proposes a practical framework for skin cancer screening research that treats image quality, lesion localization, classification, uncertainty, explanation, and external validation as connected parts of one workflow. The framework is developed through a synthesis of three supplied review papers: work on skin cancer segmentation and classification, a survey of deep learning for dermatological conditions, and a review of deep learning in cervical cancer diagnosis. The first two sources directly inform the skin imaging discussion, while the cervical cancer review contributes broader lessons about medical image pipelines and the need to validate automated screening systems. Rather than claiming new experimental results, this article presents a research proposal and evaluation plan. It recommends patient-level data separation, balanced reporting across lesion categories and skin tones, comparison with and without segmentation, calibrated uncertainty estimates, explainability checks, and external testing across imaging settings. The central argument is that clinical usefulness should be judged by robustness, fairness, interpretability, and workflow fit alongside conventional classification scores.*

Keywords: Skin cancer; deep learning; lesion segmentation; classification; explainable AI; dataset bias; external validation; computer-aided diagnosis

I. INTRODUCTION

Skin lesion assessment can involve visual examination, dermoscopy, and, when clinically indicated, further diagnostic procedures. Deep learning systems can support this process by learning image patterns associated with different lesion categories. Convolutional neural networks and encoder-decoder segmentation models are common approaches in the literature, while transfer learning and transformer-based methods are also investigated. However, a model that performs well on curated images may be less dependable when image quality, lighting, camera type, lesion appearance, or patient population changes.

The supplied reviews identify recurring concerns: limited or unevenly labelled datasets, imbalance between diagnostic classes, inadequate representation of different skin tones, high computational cost, limited model interpretability, and difficulty validating models in real clinical environments [1], [2]. These issues are related rather than independent. For example, a model trained on a narrow image distribution may appear accurate in internal testing but fail on images from another clinic. An explanation map may look persuasive while highlighting irrelevant background features. A segmentation model may improve localization for some lesions yet introduce errors when boundaries are unclear.



This article proposes a research framework that evaluates the full pathway from image input to a clinician-facing result. Its contribution is not a claim of a newly trained or clinically validated model. Instead, it provides a reproducible design for evaluating reliability, fairness, and interpretability together, drawing on the three supplied papers and identifying testable research questions.

II. RELATED LITERATURE AND RESEARCH GAP

Khalaf and colleagues review deep learning approaches to skin cancer segmentation and classification, including preprocessing, CNN-based methods, encoder-decoder architectures, transfer learning, attention mechanisms, and transformer approaches [1]. Their discussion emphasizes dataset diversity, generalization, interpretability, computational demands, and clinical integration. This suggests that model architecture alone is not enough to establish practical value.

Noronha and colleagues survey deep learning methods for a wider range of dermatological conditions and summarize datasets and methodological challenges [2]. Their review highlights the difficulty of obtaining large, accurately labelled datasets, the effects of class imbalance, and the limited variety of conditions represented in many datasets. This broader dermatology perspective matters because skin cancer models may encounter benign lesions and other visually similar conditions in real use

Youneszade and colleagues review deep learning for cervical cancer diagnosis, particularly image-based classification and segmentation in cytology and colposcopy [3]. Although cervical screening uses different image types and clinical procedures, the review provides a related medical-imaging lesson: automated systems need carefully designed preprocessing, appropriate image analysis, specialist oversight, and validation beyond headline accuracy. Its findings are used here only as methodological context, not as direct evidence about skin cancer performance.

A practical gap remains between reporting model performance on a benchmark and demonstrating dependable behavior across patient groups, institutions, and imaging conditions. This proposal addresses that gap through a staged evaluation protocol that measures classification, localization, calibration, fairness, and operational efficiency in one study design.

III. AIM AND OBJECTIVE

The aim is to propose an evaluation framework for deep learning based skin lesion analysis that prioritizes reliable performance across datasets and patient groups.

The objectives are:

1. Define a transparent image processing pipeline for lesion localization and classification.
2. Compare classification with and without a segmentation stage.
3. Measure performance by class and, where ethically and reliably recorded, by skin tone and imaging source.
4. Evaluate prediction confidence and identify cases that should be referred for expert review.
5. Examine whether explanation methods highlight clinically relevant image regions.
6. assess external generalization and computational requirements before considering clinical use.

IV. PROPOSED FRAMEWORK

4.1 Data collection and governance.

Use legally accessible, appropriately licensed dermoscopic image datasets with documented labels and metadata. Record the source, acquisition conditions, label provenance, and known limitations. Remove duplicates and split data at patient level so that images from the same patient cannot appear in both training and test partitions. Keep a final external dataset untouched until model selection is complete. Use of skin tone or demographic attributes should follow ethical approval, data minimization, and appropriate governance.ayered architecture of the RAG based Legal Document Assistant.



4.2 Preprocessing.

Apply consistent resizing and intensity normalization while preserving lesion detail. Detect or reduce common artefacts such as hair, rulers, pen marks, and uneven illumination only when the operation can be applied consistently. Document every preprocessing step and evaluate whether it removes diagnostic information. Augmentation may include moderate rotations, flips, crops, and lighting variation, but transformations should not create medically implausible examples or distort lesion boundaries.

4.3 Lesion localization.

Evaluate a segmentation model, such as a U-Net family architecture, to estimate the lesion region. Compare the segmentation output against expert masks when available. The purpose is to test whether isolating the lesion improves classification, not to assume that segmentation always helps. A parallel baseline should classify the original image so the impact of segmentation can be measured directly

4.4 Classification

Train a compact CNN baseline and compare it with a transfer learning model. If resources permit, include a transformer or hybrid model as a secondary comparison. Use class-weighted loss or carefully controlled resampling to address class imbalance. Keep model selection separate from final testing. Report per-class sensitivity, specificity, precision, F1 score, balanced accuracy, and area under the receiver operating characteristic curve where appropriate. Overall accuracy should never be the only reported measure.

4.5 Confidence and referral

Add a calibration stage using a validation set, and measure whether predicted confidence corresponds to observed correctness. Define a referral rule for uncertain predictions, unusual inputs, or images outside the training distribution. The system should be allowed to abstain rather than forcing a confident label for every image. Referral thresholds must be selected with clinical input and evaluated for their effect on missed cases and workload.

4.6 Explanation review

Apply an explanation method, for example Grad-CAM, to visualize regions that influenced model output. Explanation maps should be assessed with structured checks, including whether highlighted regions overlap the lesion and whether predictions change when irrelevant background areas are altered. Visual explanations are aids for inspection, not proof that the model reasons correctly.

4.7 Output and human oversight

The interface should present the predicted category, confidence or uncertainty status, a localization overlay when available, and a clear statement that the output is decision support rather than a diagnosis. Clinical decisions remain with qualified professionals. The system should retain audit information needed to investigate errors, subject to privacy and security requirements

V. RESEARCH QUESTION AND HYPOTHESES

RQ1: Does adding lesion segmentation improve classification performance compared with classification from the full image? Hypothesis: benefits will vary by dataset and lesion type, so the effect should be measured rather than assumed

RQ3: Do results vary across diagnostic classes, skin tones, and image sources? Hypothesis: aggregate scores may conceal meaningful subgroup differences, especially where training examples are sparse

RQ4: Does confidence based referral reduce high confidence errors? Hypothesis: a calibrated abstention strategy may improve the reliability of accepted predictions, while increasing the number of cases requiring expert review.



RQ4: Does confidence based referral reduce high confidence errors? Hypothesis: a calibrated abstention strategy may improve the reliability of accepted predictions, while increasing the number of cases requiring expert review.

VI. EVALUATION METHODOLOGY

The study should use three partitions: training, validation, and a locked internal test set, with patient-level separation. A second institution or independently collected dataset should be reserved for external testing. If multiple public datasets are combined, dataset identity should be tracked and leakage checks should be performed. Label definitions should be harmonized carefully; labels that are not clinically equivalent should not be merged simply to increase sample size.

The baseline comparison should include: (A) image classification without segmentation, (B) segmentation followed by classification, and (C) a model with confidence based referral. Each model should be trained using the same split and comparable tuning budgets. Report confidence intervals, preferably using patient-level bootstrap resampling, and disclose the number of cases in every class and subgroup. For imbalanced data, report macro-averaged metrics and confusion matrices as well as class-specific sensitivity and specificity.

Segmentation can be evaluated with Dice coefficient and intersection over union when expert masks exist. Classification can be assessed using balanced accuracy, macro F1, sensitivity, specificity, and AUROC, while precision-recall analysis is useful for rare classes. Calibration can be measured with reliability diagrams and expected calibration error. Operational evaluation should record model size, inference time, memory use, and the proportion of cases referred for human review. Results should be interpreted in the context of the dataset and not generalized to all patients without supporting evidence

VII. EXPECTED CONTRIBUTION

The proposed work is expected to contribute a structured evaluation protocol rather than an unsupported claim of improved diagnostic accuracy. First, it will make the value of segmentation measurable by comparing otherwise similar pipelines. Second, it will expose subgroup weaknesses that aggregate metrics can conceal. Third, uncertainty and referral handling will be evaluated as part of system performance. Fourth, explanation checks and external validation will help identify failure modes before any deployment is considered. Finally, reporting computational requirements can help determine whether a model is practical for clinics with limited resources.

These are expected contributions and research hypotheses. They must be confirmed through experiments; no performance numbers are asserted in this draft.

VIII. LIMITATIONS AND ETHICAL CONSIDERATIONS

This proposal does not include a newly trained model, patient data, or experimental results. Its value is therefore methodological, and the framework requires implementation and validation before conclusions about performance can be drawn. Public datasets may have incomplete demographic information, inconsistent labels, duplicate images, and limited representation of some lesion categories. Skin tone metadata may be unavailable or measured inconsistently, so subgroup analysis must not rely on unsupported assumptions

Medical images are sensitive data. Researchers should follow dataset licences, institutional requirements, and applicable privacy rules. Synthetic images should not be treated as a replacement for representative real clinical data. A model can support triage or clinician review but should not be presented as a stand-alone diagnostic authority. Prospective studies, workflow evaluation, and regulatory review may be required before real-world use

IX. FUTURE SCOPE

Future research can examine multimodal systems that combine images with carefully selected clinical context, while testing whether additional information genuinely improves performance. Lightweight models may support resource-constrained environments, but compression must be evaluated for its effect on rare classes and calibration. Federated learning may be studied where data cannot be centralized, provided privacy protections and cross-site evaluation are



explicit. Further work should also investigate temporal monitoring for model drift, standardized reporting of subgroup performance, and clinician-centered studies measuring whether the tool improves decisions without increasing unnecessary procedures

X. CONCIUSION

Deep learning offers useful methods for skin lesion segmentation and classification, but accuracy on a single dataset is not sufficient evidence of clinical readiness. The reviewed literature points to recurring challenges in dataset quality, imbalance, demographic representation, interpretability, generalization, and computational cost [1], [2]. Related work on cervical cancer image analysis reinforces the broader need for carefully designed medical imaging pipelines and validation [3]. This paper proposes a testable framework that combines segmentation comparison, classification, calibration, explanation review, subgroup analysis, external testing, and human referral. The framework deliberately avoids claiming results that have not been measured. Its core recommendation is to evaluate a skin cancer screening model not only by how often it is correct in a benchmark, but also by where it fails, how uncertain it is, whom it serves less well, and whether clinicians can use its output safely

REFERENCES

- [1] A. D. Khalaf, H. Hamdan, A. A. Halin, and N. Manshor, "Segmentation and Classification of Skin Cancer Diseases Based on Deep Learning: Challenges and Future Directions," *IEEE Access*, vol. 13, pp. 90163–90184, 2025, doi: 10.1109/ACCESS.2025.3569170
- [2] S. S. Noronha, M. A. Mehta, D. Garg, K. Kotecha, and A. Abraham, "Deep Learning-Based Dermatological Condition Detection: A Systematic Review With Recent Methods, Datasets, Challenges, and Future Directions," *IEEE Access*, vol. 11, pp. 140348–140381, 2023, doi: 10.1109/ACCESS.2023.3339635.
- [3] N. Youneszade, M. Marjani, and C. P. Pei, "Deep Learning in Cervical Cancer Diagnosis: Architecture, Opportunities, and Open Research Challenges," *IEEE Access*, vol. 11, pp. 6133–6150, 2023, doi: 10.1109/ACCESS.2023.3235833

