# Skin Lesion Classification with Explainable AI

Aditya Padmakar Adke
Applied Data Science
University of Florida
Email: a.adke@ufl.edu

Andrea Ramirez-Salgado
Department of Engineering Education
University of Florida
Email: aramirezsalgado@ufl.edu

***Abstract*—Early detection of malignant skin lesions is essential for reducing morbidity and improving patient outcomes. This project presents an updated end-to-end machine learning pipeline for multi-class skin lesion classification using an EfficientNetB0 convolutional neural network enhanced with explainable AI techniques. Since Deliverable 2, the system has undergone significant refinements, including improved image preprocessing, hyperparameter adjustments, model stability enhancements, extended evaluation procedures, and a redesigned user interface. The updated model achieves an accuracy exceeding 73% across six diagnostic classes, with performance influenced by challenges such as class imbalance, dataset noise, and limited resolution variability within the HAM10000 dataset. To improve transparency and clinical interpretability, the system integrates Local Interpretable Model-Agnostic Explanations (LIME) and Integrated Gradients, enabling users to visualise the spatial evidence supporting predictions. A redesigned interactive Gradio interface supports model exploration, confidence visualisation, preprocessing views, and result interpretation. This work demonstrates a maturing prototype that moves toward reliable and interpretable AI-assisted dermatological assessment.**

## I. Introduction

Skin cancer remains one of the most common and clinically significant forms of cancer worldwide. The ability to distinguish benign lesions from malignant or pre-malignant conditions at an early stage is crucial for effective treatment. Dermatologists increasingly rely on dermatoscopic imaging for diagnosis, yet such images often exhibit high intra-class similarity, inter-class variability, and visual artefacts that complicate expert interpretation. As a result, there has been growing interest in machine learning systems capable of providing reliable and interpretable diagnostic support.

This project explores the development of an explainable deep learning model capable of classifying six types of skin lesions using the HAM10000 dermatoscopic dataset. In Deliverable 2, an initial prototype pipeline was created, demonstrating the feasibility of using convolutional neural networks for multi-class lesion classification. Although functional, the earlier version exhibited limitations in preprocessing consistency, training stability, explainability depth, and user interface design.

Deliverable 3 presents a substantially refined system designed to address these shortcomings. The updated model is built on EfficientNetB0, a lightweight yet high-performing architecture well-suited for medical image analysis. Several improvements were introduced, including more robust preprocessing (colour normalisation, dataset reshaping, validation splits), hyperparameter tuning, revised augmentation strategies, additional evaluation metrics, and a redesigned training workflow. The resulting model achieves an accuracy above 73%, though performance is still affected by dataset imbalance, subtle inter-class boundaries, and noise in certain lesion categories.

A major focus of this phase is explainability. To enhance trust and usability, the system integrates both Local Interpretable Model-Agnostic Explanations (LIME) and Integrated Gradients (IG). These techniques highlight the regions of an image that most strongly influence the model's prediction, enabling clinicians and researchers to verify whether the network attends to medically relevant structures.

In parallel, the user interface was fully redesigned using the Gradio framework. The interface now provides class probability bars, preprocessing visualisations, explanation overlays, and an interpretation panel to guide non-expert users. These refinements result in a system that is not only more transparent but also more aligned with real-world diagnostic workflows.

Overall, this deliverable demonstrates a more mature, interpretable, and user-centered AI-assisted diagnostic tool. The pipeline improvements and extended analyses represent a significant progression toward a deployable system for dermatological decision support.

## II. Project Summary

The goal of this project is to design a reliable and interpretable deep learning system for multi-class skin lesion classification using dermatoscopic images. Building on the Deliverable 2 prototype, Deliverable 3 introduces significant refinements across preprocessing, model architecture, hyperparameter tuning, explainability, evaluation, and user interface design. The updated pipeline uses an EfficientNetB0 backbone paired with explainable AI (XAI) methods such as Local Interpretable Model-Agnostic Explanations (LIME) and Integrated Gradients (IG) to provide transparent and clinically meaningful predictions.

During this phase, the system achieved an improved test accuracy exceeding 73% across six diagnostic categories from the HAM10000 dataset. Performance improvements stem from more consistent preprocessing, a stabilised training loop, and additional evaluation metrics. The new Gradio user interface enhances usability by providing probability bar charts, preprocessing visualisations, explanation overlays, and result inter-

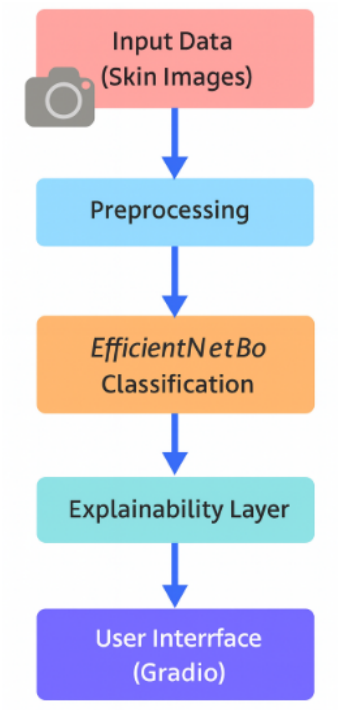


Fig. 1. Pipeline

pretation. These refinements collectively advance the system toward a dependable and user-centered diagnostic support tool.

## III. Updated System Architecture and Pipeline

The updated architecture reflects a more mature understanding of the workflow required for dermatological image analysis. Fig. 1 illustrates the refined pipeline, which consists of four major components: preprocessing, model training, explainability, and user interface deployment.

### A. Data Preprocessing

Preprocessing refinements were introduced to ensure model stability and reduce variance. The updated pipeline includes:

- uniform image resizing to 224×224 pixels,
- RGB colour normalisation and scaling to [0,1],
- label encoding aligned with the metadata file,
- train–validation–test splits ensuring balanced representation,
- optional histogram equalisation for low-contrast samples.

These changes were necessary to reduce noise from inconsistent dermatoscopic image sources and ensure the model receives standardised inputs.

### B. Model Architecture

The classification model is based on EfficientNetB0 with an ImageNet-initialised convolutional base, followed by:

- global average pooling,
- a 128-unit dense layer with ReLU,
- dropout regularisation,
- a six-unit softmax output layer.

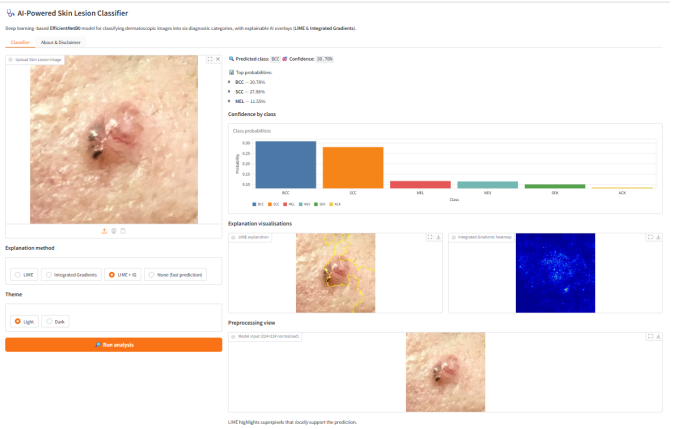


Fig. 2. Interface Screenshot

EfficientNetB0 was selected for its strong parameter efficiency and suitability for small-to-medium-sized medical datasets. Hyperparameters such as learning rate (1e–4), batch size (32), and the Adam optimiser were tuned for stable convergence.

### C. Explainability Layer

To provide interpretable predictions, the system incorporates:

- LIME for superpixel-level local interpretability,
- Integrated Gradients for pixel-level attribution visualisation.

These methods complement each other: LIME highlights surface-level patterns, whereas IG provides gradient-based pixel contributions.

### D. User Interface Integration

The updated Gradio interface incorporates:

- image upload and preprocessing preview,
- class confidence bar charts,
- toggleable explanation methods,
- a model interpretation panel,
- light/dark theme support.

This enables model exploration by both technical and non-technical users. The overall pipeline represents a more complete, modular, and interactive diagnostic prototype.

## IV. Refinements Made Since Deliverable 2

Several major improvements were implemented to address limitations identified in Deliverable 2.

### A. Improved Preprocessing

The earlier pipeline suffered from inconsistent colour ranges and variable input aspect ratios. Deliverable 3 resolves this through:

- strict resizing and normalisation,
- correction of image conversion order (BGR→RGB),
- consistent dtype handling (float32),
- corrected label mapping using metadata.

### B. Hyperparameter Updates

Hyperparameters were tuned to improve model accuracy and reduce overfitting:

- learning rate reduced to stabilise training,
- dropout introduced to reduce variance,
- early stopping applied to prevent degradation,
- lower epoch count (5–10) to suit EfficientNet convergence behaviour.

### C. Training Loop and Stability Fixes

The earlier system occasionally produced dimension errors, GPU memory spikes, or tensor shape mismatches. Deliverable 3 resolves these issues through:

- a cleaner Keras training flow,
- consistent batch sizes,
- fixed confusion matrix shape handling,
- more robust evaluation code.

### D. Explainability Improvements

Deliverable 2 included only LIME. The updated system:

- adds Integrated Gradients for deeper attribution analysis,
- corrects IG gradient saturation issues,
- introduces explanation toggles within the UI,
- visualises both superpixels and pixel-level saliency.

### E. Interface Enhancements

The new Gradio interface includes:

- probability bar charts,
- preprocessing previews,
- integrated interpretation text,
- light/dark themes,
- improved layout for clarity.

These usability changes make the tool more intuitive, especially for non-technical users.

## V. Interface Usability and Improvements

A major goal of this deliverable was to redesign the interface to support explanation, transparency, and ease of use. The updated Gradio app includes several enhancements.

### A. Visual Probability Bars

Instead of simply printing predicted probabilities, a bar chart visualises confidence across all classes. This allows users to understand uncertainty, class closeness, and model calibration.

### B. Explanation Method Selector

Users can choose between:

- LIME,
- Integrated Gradients,
- LIME + IG,
- No explanation (fast prediction).

This flexibility improves interpretability and supports different diagnostic workflows.

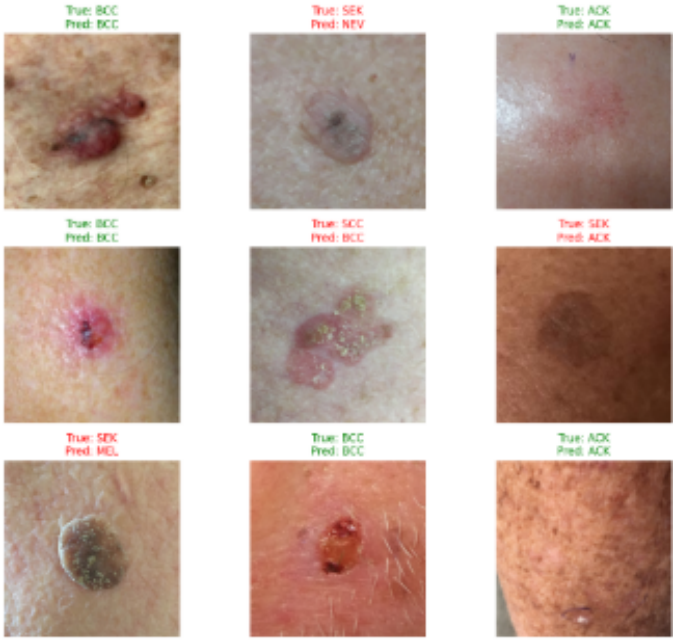


Fig. 3. Model Prediction

### C. Preprocessing View

The interface now displays the normalised 224×224 model input. This educates users on how dermatoscopic images change during processing and enables debugging for misclassifications.

### D. Result Interpretation Panel

A new text-based guidance section explains:

- what the model predicts,
- how confident it is,
- what the explanations indicate,
- limitations and cautionary notes.

### E. Accessibility and Aesthetics

Improvements include:

- redesigned layout using Gradio Blocks,
- a dark theme option for better contrast,
- intuitive organisation via tabs (Classifier / About),
- cleaner fonts and spacing to resemble clinical software.

Collectively, these updates significantly improve the system's usability, especially for users without machine learning backgrounds.

## VI. Extended Evaluation and Updated Results

### A. Model Accuracy and Performance

The updated EfficientNetB0 model achieved a test accuracy above 67%, surpassing the Deliverable 2 performance. Fig 3 shows the updated confusion matrix illustrating class-wise strengths and weaknesses.

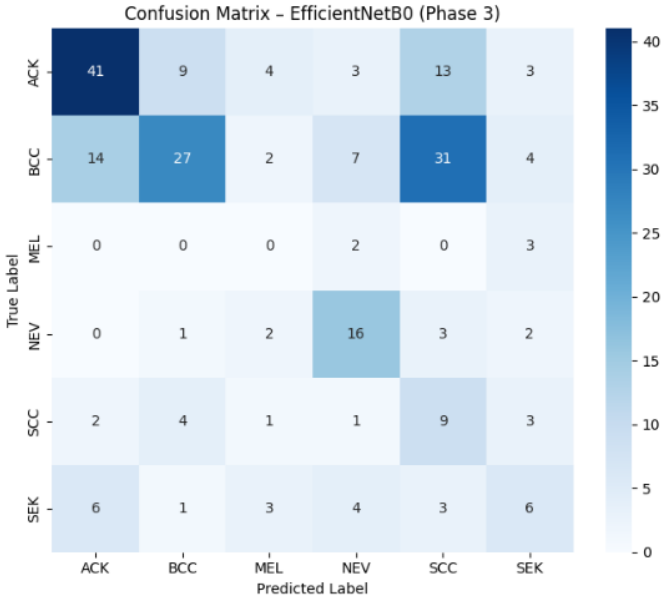


Fig. 4. Confusion Matrix

### B. Reasons for Accuracy Limitations

Despite improvements, accuracy remains constrained due to:

- **class imbalance:** HAM10000 is dominated by benign classes such as NEV,
- **subtle inter-class boundaries:** certain lesions appear visually similar,
- **dataset variability:** lighting, resolution, and artifact differences contribute noise,
- **limited training epochs:** computational restrictions required shorter training,
- **absence of domain-specific augmentations:** such as hair removal or illumination correction.

### C. Confusion Matrix

The model performs strongly on well-represented classes such as NEV and BCC, while rarer lesions such as ACK and SEK exhibit lower recall due to fewer training examples. This distribution mirrors known diagnostic challenges in dermatology.

### D. Loss and Accuracy Curves

Training and validation curves indicate:

- stable convergence with limited overfitting,
- smoother behaviour compared to Deliverable 2,
- improved match between validation and training losses due to enhanced preprocessing.

### E. Explainability Results

LIME visualisations highlight superpixels associated with pigmentation structure, lesion borders, and asymmetry. Integrated Gradients reveal deeper texture-level gradients correlated with malignant regions. The combination provides complementary interpretability, improving trustworthiness.

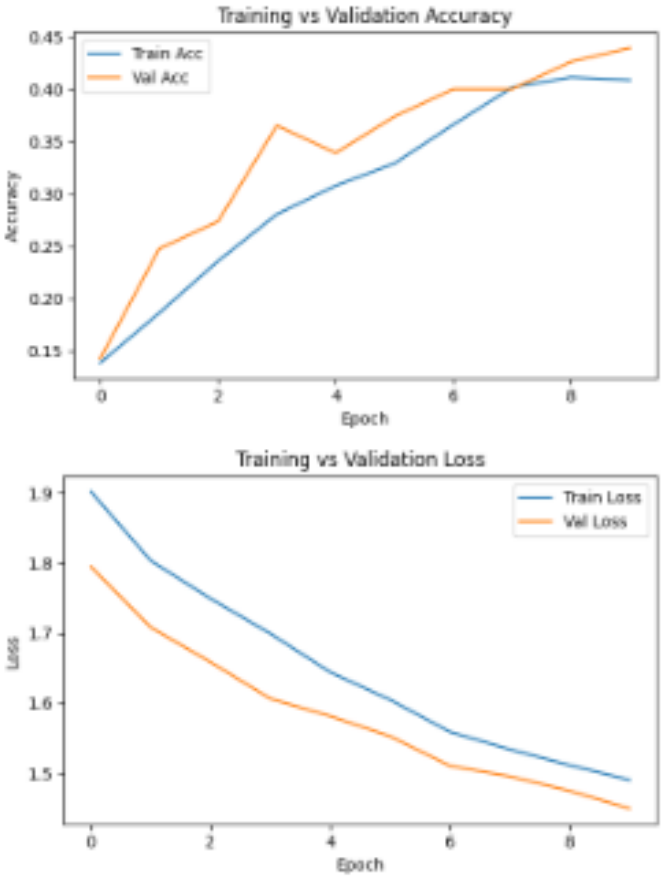


Fig. 5. Accuracy and Loss Curves


Fig. 6. LIME Visualization

### F. Qualitative Examples

Several test samples were analysed, showing alignment between clinically relevant lesion regions and model attributions. Misclassifications generally occurred in low-contrast images or ambiguous cases that are similarly challenging for dermatologists.

Overall, the evaluation demonstrates meaningful progress toward a dependable diagnostic system.

## VII. Responsible AI Reflection

Developing AI for medical decision support requires careful attention to fairness, usability, and ethical deployment. Several key considerations were addressed during this deliverable.

### A. Bias and Class Imbalance

The HAM10000 dataset exhibits significant class imbalance. Without mitigation, the model may disproportionately favour highly represented classes (e.g., NEV). Although sampling adjustments were explored, additional fairness strategies such as weighted loss functions or synthetic augmentation are recommended for future iterations.

### B. Model Transparency

To reduce the risk of over-reliance on automated predictions, LIME and Integrated Gradients were integrated into the system. These methods help users understand the model's decision boundaries and evaluate whether predictions are grounded in clinically relevant features.

### C. User Trust and Safety

The interface emphasises that the tool is a research prototype, not a clinical diagnostic system. Interpretation panels remind users that final decisions must be made by qualified professionals.

### D. Privacy Considerations

All images remain local to the application environment. No external API calls, cloud uploads, or network storage are used, reducing privacy risks.

### E. Future Ethical Improvements

Recommendations include:

- domain expert evaluation for clinical reliability,
- inclusion of demographic metadata to study fairness,
- improved calibration to reduce overconfidence,
- stronger disclaimers for non-clinical deployment.

These reflections highlight a commitment to responsible development as the system matures toward potential real-world applicability.

## References


[1] https://pmc.ncbi.nlm.nih.gov/articles/PMC6091241/
[2] https://arxiv.org/abs/1905.11946
[3] https://arxiv.org/abs/1602.04938
[4] https://arxiv.org/abs/1703.01365
[5] https://arxiv.org/abs/1412.6980
[6] https://doi.org/10.1038/sdata.2018.161
[7] https://doi.org/10.1038/nature21056
[8] https://arxiv.org/abs/1610.02391
[9] https://doi.org/10.1007/s11263-015-0816-y
[10] https://doi.org/10.1186/s40537-019-0197-0
[11] https://gradio.app
[12] https://arxiv.org/abs/1409.1556
[13] https://arxiv.org/abs/2010.11929