# Skin Lesion Classification with Explainable Deep Learning

KEREM DOGAN, Georgia Southern University, USA

This study presents a deep learning framework for multi-class skin lesion classification using dermoscopic images, combined with post-hoc explainability techniques to improve model transparency. The dataset, derived from the ISIC archive, contains ten heterogeneous lesion categories with substantial class imbalance and visual variability. An EfficientNet-based classifier was trained through a two-stage process involving frozen feature extraction and fine-tuning, followed by evaluation using accuracy, F1-score, ROC curves and confusion matrices. The model achieved moderate overall performance, with strong discriminative ability for several classes but reduced precision on rare categories. To interpret model decisions, Grad-CAM and Grad-CAM++ were applied to visualize class-relevant regions. Correct predictions typically aligned with clinically meaningful structures, whereas misclassifications often resulted from attention shifts toward artefacts or background regions. These findings demonstrate that explainability tools provide essential insight into model behavior beyond conventional metrics, revealing both clinically plausible and erroneous decision patterns. The combined analysis highlights the potential of deep learning for dermatological decision support while emphasizing the need for improved data balance, robustness strategies and explainable model design in future work.

Additional Key Words and Phrases: Skin lesion classification, Dermoscopy, Deep learning, EfficientNet, Explainable AI, Grad-CAM, Grad-CAM++, Medical image analysis, Interpretability, Error analysis.

**ACM Reference Format:**
Kerem Dogan. 2025. Skin Lesion Classification with Explainable Deep Learning. 1, 1 (December 2025), 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 Introduction

Skin lesions encompass a wide spectrum of benign and malignant abnormalities that arise from diverse pathological processes, and their visual presentation often varies considerably across patients and imaging conditions. Dermoscopic analysis plays a central role in the early detection of malignant lesions, yet even experienced clinicians encounter substantial diagnostic uncertainty when differentiating visually similar categories. Subtle variations in pigmentation patterns, irregular structures, and heterogeneous backgrounds make the assessment challenging, and misclassification can have severe clinical implications since survival rates for melanoma are strongly tied to timely and accurate diagnosis. These difficulties have motivated a sustained interest in computational systems that can support clinical judgement by providing consistent, reproducible assessments of lesion images.

The emergence of large-scale image datasets and modern deep learning architectures has created new possibilities for automating such diagnostic tasks. Earlier approaches to computer-aided dermatology relied on hand-crafted features engineered from color, texture, or structural cues, and their performance was limited by the designer's ability to anticipate relevant visual patterns. Deep learning techniques reshaped this landscape by introducing end-to-end representation learning, where convolutional and related neural architectures learn hierarchical features directly from raw imagery. [1] Ekundayo and Ezugwu (2025) trace the conceptual development of these models, emphasizing how advances in network depth, training paradigms, and computational capabilities transformed neural networks into powerful tools for high-dimensional perception tasks, including medical imaging. Their discussion provides a useful foundation for understanding why deep models are well suited to complex visual domains like skin lesion classification, where subtle distinctions must be captured within highly variable inputs.

Research across numerous imaging disciplines reinforces this suitability. Reviews such as [2] Alzubaidi et al. (2021) highlight how convolutional and hybrid architectures have consistently surpassed traditional pipelines, particularly in settings where data complexity challenges manual feature engineering. These models leverage large annotated corpora, regularization strategies, and transfer learning to achieve robust performance, yet they are also sensitive to issues that frequently arise in dermatology datasets, including class imbalance, limited samples for rare lesion types, and visually ambiguous boundaries. Comparable observations

Author's Contact Information: Kerem Dogan, Georgia Southern University, Statesboro, Georgia, USA, kd20511@georgiasouthern.edu.


Manuscript submitted to ACM

appear in domains outside medicine as well. [3] Li et al. (2018) demonstrate how deep models outperform conventional methods in remote sensing classification, a task characterized by heterogeneous backgrounds, fine-grained inter-class differences, and noise introduced by acquisition conditions. These characteristics mirror the challenges of dermoscopic images and further justify the reliance on learned feature hierarchies for improved generalization and diagnostic consistency.

Despite these advantages, the opacity of deep learning models presents a notable barrier to their clinical adoption. Dermatological diagnosis depends not only on accurate predictions but also on an articulated reasoning process that clinicians can scrutinize, question, and integrate with their own expertise. Explainability therefore becomes a necessary complement to performance. Work in explainable artificial intelligence has offered conceptual and methodological tools for addressing this need. [4] Love et al. (2023) outline major families of explanation techniques and discuss how concepts such as fidelity, human interpretability, stability, and fairness shape the design of transparent machine learning systems in high-stakes environments. Their framing underscores a central tension in medical AI: powerful models offer strong predictive capabilities but often operate in ways that remain inaccessible to human users.

In the context of visual data, explanations must bridge low-level image processing operations and the higher-level concepts that guide diagnostic reasoning. [5] Singh et al. (2020) show how interpretability can be embedded throughout computer vision pipelines, employing mechanisms like feature attribution maps, attention-based visualization, and prototype-based reasoning to reveal which image regions or structures most influence model decisions. Such approaches are particularly valuable for dermoscopic imagery, where clinically meaningful cues such as irregular pigment networks, asymmetries, streaks, and local color transitions, must be identifiable within explanations. [6] Dharshini et al. (2023) further provide a taxonomy that distinguishes between intrinsically interpretable models and post-hoc methods designed for deep networks, detailing the trade-offs between transparency, performance, computational cost, and local versus global insights. Their analysis clarifies why post-hoc visual explanation tools are a practical choice for convolutional classifiers in healthcare settings, where model complexity is typically unavoidable but interpretability remains indispensable.

Against this backdrop, the present study develops a deep learning–based skin lesion classification system enriched with explainability mechanisms aimed at producing clinically interpretable visual rationales. The project seeks to combine competitive predictive performance with transparent, inspection-friendly explanations that can help clinicians evaluate the plausibility of model outputs. By integrating established neural architectures with post-hoc visual interpretation tools and assessing their alignment with domain-relevant lesion characteristics, this work aims to contribute a small but meaningful step toward AI systems that are not only accurate but also trustworthy and usable in dermatological practice.

## 2 Literature Review

The use of machine learning and deep learning in medical image analysis has expanded rapidly as clinicians seek computational tools capable of assisting with complex diagnostic tasks. Early research in fields such as neuro-oncology illustrates both the opportunities and the limitations of automated image classification. [7] Solanki et al. (2023) compare classical radiomic pipelines with deep learning methods for brain tumor categorization, demonstrating that convolutional networks trained end to end generally surpass traditional feature-based approaches when sufficient data augmentation and regularization are applied. Their results highlight recurring challenges in medical imaging such as limited annotated datasets, imbalance between diagnostic categories and the need for models whose predictions can be scrutinized by clinicians rather than accepted as opaque outputs. Similar themes appear across broader oncology imaging research. [8] Hosny et al. (2024) explore radiomics and deep learning for cancer assessment and show that learned feature hierarchies often produce more discriminative biomarkers than manually engineered descriptors, yet they caution that limited interpretability, annotation inconsistencies and variability in imaging protocols remain major barriers to clinical integration. These findings provide a conceptual backdrop for skin lesion analysis, where the visual heterogeneity of dermoscopic images and the clinical risks associated with misclassification similarly require models that pair strong performance with transparent reasoning.

As deep learning systems became established in clinical imaging workflows, researchers increasingly focused on enhancing their interpretability so that their internal reasoning could be understood by non-technical experts. [9] Vermani et al. (2025) present an explainable deep learning framework for lung disease diagnosis that integrates high-performing neural networks with interactive visual and textual explanations. Their study demonstrates that meaningful explanations help clinicians verify that model outputs align with established radiological patterns, detect potential biases and identify misclassifications that would otherwise remain

hidden. Trust and adoption improved when explanations were consistent, faithful to the underlying model and expressed in terms familiar to radiologists. This observation has direct relevance for skin lesion classification; dermatologists similarly rely on recognizable visual cues, and an AI system that highlights relevant structures and pigmentation features can support more confident decision-making while mitigating the risks associated with opaque predictions.

Building on developments in general medical imaging, a substantial body of work has examined the application of deep learning to dermoscopic images specifically. [10] Arshad et al. (2025) analyse a pipeline that incorporates preprocessing, lesion-focused feature extraction and fine-tuning of pre-trained architectures to differentiate benign from malignant lesions. Their findings confirm that convolutional models typically outperform classical methods relying on hand-crafted features, although challenges such as dataset bias, imaging variability and the difficulty of generalizing across populations remain unresolved. [11] Ahmad et al. (2023) reach similar conclusions, proposing a deep learning framework enriched with segmentation, colour normalization and augmentation steps to mitigate the instability introduced by heterogeneous acquisition conditions. While their models achieve strong benchmark performance, the authors note that reliability and interpretability continue to hinder clinical deployment. [12] Jahan et al. (2024) likewise demonstrate that deep neural networks can outperform traditional baselines but emphasize that issues such as overfitting, class imbalance and the gap between experimental benchmarks and real-world scenarios must be addressed through more principled validation and interpretability assessments. Collectively, these studies establish the technical viability of deep learning for skin lesion classification while underscoring that accuracy alone does not guarantee clinical readiness.

Research that combines deep learning with explicit explainability mechanisms provides the closest precedent for the objectives of this thesis. [13] Nigar et al. (2022) propose a CNN-based system that incorporates post-hoc saliency and class-activation-based visualizations to highlight the regions of dermoscopic images that drive model predictions. Their approach demonstrates that high predictive performance can coexist with interpretable explanations that dermatologists can qualitatively assess for plausibility, enabling the detection of spurious correlations, improving reliability and fostering trust in AI-supported diagnosis. While their contributions are significant, questions remain about explanation consistency, robustness and how well these visual cues align with clinically meaningful features across diverse lesion types. These gaps motivate further refinement of explainable deep learning approaches, particularly methods that provide clearer, more stable and more clinically grounded visual rationales.

## 3 Methodology

The methodological framework of this study was designed to support the development of a deep learning pipeline that performs accurate skin lesion classification while simultaneously producing interpretable visual explanations. The system integrates several sequential components, beginning with dataset preparation and continuing through model construction, training, evaluation and post-hoc interpretability analysis. This staged workflow allows each component to be optimized independently while ensuring that the final model behaves consistently across both predictive and explanatory dimensions. Central to the overall design is the recognition that classification performance alone is insufficient for clinical deployment; therefore, explainability techniques are embedded into the pipeline from the outset, guiding architectural choices and evaluation procedures.

The framework proceeds in four major steps. First, dermoscopic images are preprocessed and standardized to mitigate variations in resolution, illumination and acquisition conditions, and to ensure compatibility with modern convolutional architectures. Second, a deep learning model is trained using transfer learning and fine-tuning strategies, enabling the network to capture subtle visual distinctions between lesion types despite limited dataset size. Third, the trained model is evaluated using a combination of classification metrics and error-focused analyses to diagnose biases, identify failure modes and understand class-specific performance characteristics. Finally, explainability techniques, primarily Grad-CAM and its extension Grad-CAM++, are applied to visualize the internal reasoning of the model. These methods highlight the discriminative image regions that contribute to each prediction, allowing clinicians and researchers to inspect whether the model relies on plausible dermatological structures rather than irrelevant artefacts or background patterns.

Taken together, this multi-stage methodological structure provides a transparent and reproducible framework for skin lesion analysis. Each phase is designed to contribute a specific layer of functionality such as data quality control, feature learning, diagnostic evaluation and interpretability, resulting in a system that not only classifies lesions but also exposes the underlying evidence supporting each decision. Such a framework aligns closely with contemporary expectations for trustworthy medical AI systems, where explainability is not treated as an add-on but as an integral component of the model development lifecycle.

### 3.1 Research Questions

(1) To what extent can a convolutional neural network trained on dermoscopic images achieve accurate multi-class skin lesion classification across heterogeneous diagnostic categories?
(2) How do different lesion categories affect the classifier's performance, and what do error patterns reveal about biases, dataset imbalance and structural limitations of the model?
(3) Can post-hoc visual explanation techniques such as Grad-CAM and Grad-CAM++ produce faithful and stable localization maps that meaningfully highlight the structures responsible for the model's predictions?
(4) Does the integration of explainability methods improve the transparency and trustworthiness of the classifier, particularly in cases where the model is uncertain or incorrect?

### 3.2 Dataset

The experiments were conducted on the ISIC Barcelona (BCN20000) dermoscopic image collection, which provides a large set of annotated skin lesion images curated for computer-aided diagnosis research. The metadata file released with the collection includes, for each image, a unique identifier (isic_id), approximate patient age, sex, several fields describing the anatomical site, lesion-level identifiers, diagnostic labels at different levels of granularity (diagnosis_1, diagnosis_2, diagnosis_3), confirmation type, an indicator of whether the lesion is melanocytic, and the imaging modality (image_type). For this study, the diagnosis_3 field was used as the primary ground-truth label, as it encodes the clinically meaningful lesion type, while isic_id and image_type were used to construct valid file paths and to restrict the dataset to dermoscopic images only.
Starting from the full metadata table, entries without a valid diagnosis_3 label were removed, and only dermoscopic images were retained. After these filters, the working dataset contained 17,639 images, each linked to a single lesion type. The diagnosis_3 labels span ten diagnostic categories, including Nevus (5,647 images), Melanoma (4,003), Basal cell carcinoma (3,676), Seborrheic keratosis (1,268), Solar or actinic keratosis (1,088), Melanoma metastasis (633), Squamous cell carcinoma (559), Scar (314), Solar lentigo (283), and Dermatofibroma (168). This distribution exhibits a clear class imbalance, with common lesions such as nevi heavily over-represented compared with rarer entities like dermatofibroma or melanoma metastasis. To make the data suitable for model training, the string labels from diagnosis_3 were mapped to integer indices, yielding a ten-class classification problem where each image belongs to exactly one lesion category.
Only the image data and the diagnosis_3 labels were used as inputs to the deep learning model; demographic and anatomical metadata (age, sex, anatomical site) were not incorporated into the training pipeline, in order to focus the analysis on image-based visual cues. For each isic_id, an absolute file path was constructed to the corresponding JPEG file, and any entries whose image file could not be located on disk were discarded. The resulting dataset was then partitioned into training, validation and test subsets using stratified splits based on the encoded class labels. Approximately 15% of the images were reserved as a held-out test set, while the remaining data were further divided into a training set and a validation set, with about 72% of the images used for training, 13% for validation and 15% for testing overall. All images were later resized to 224×224 pixels and normalized within the data pipeline to match the input requirements of the EfficientNet-based classifier described in the subsequent sections.

### 3.3 Phase 1: Data Preprocessing and Standardization

The first phase of the methodology focuses on preparing dermoscopic images for downstream learning tasks by establishing a standardized and reproducible preprocessing pipeline. Given the substantial variability in image resolution, acquisition settings and illumination conditions across the ISIC dataset, systematic preprocessing is essential to ensure that the deep learning model receives consistent and diagnostically meaningful inputs. All raw images were processed through a sequence of operations implemented using TensorFlow's tf.data framework, which enables efficient, parallelized loading and transformation of large-scale image collections.
Each image was first converted to a fixed spatial resolution of 224 × 224 pixels, aligning with the input requirements of EfficientNet-based models and ensuring uniformity across batches. Images were then cast to floating-point format and rescaled from the original 0–255 pixel range to values in either [0, 1] or the standardized ImageNet normalization range, depending on the phase of training. This normalization step stabilizes gradient magnitudes during backpropagation and reduces sensitivity to illumination differences inherent in dermoscopic photography. Since the ISIC dataset includes only dermoscopic images, no filtering by modality

was necessary after path construction, but file integrity checks were applied to ensure that all entries in the metadata table corresponded to accessible image files.

To improve generalization and mitigate overfitting, data augmentation was applied on-the-fly during training. Augmentation included random horizontal and vertical flips, rotations, zoom transformations and variations in brightness or contrast, each selected to simulate plausible dermatological imaging conditions while preserving clinically relevant structures. These transformations help the model become robust to small variations in viewpoint and lesion orientation, which are common in dermoscopy. Augmentation was not applied to the validation or test sets to ensure unbiased evaluation. All preprocessing operations were encapsulated within the TensorFlow data pipeline, enabling images to be decoded, transformed, batched and prefetched efficiently during model training.

By enforcing consistent spatial resolution, intensity normalization and principled augmentation, this preprocessing phase provides a stable foundation for the subsequent learning stages. The standardized inputs allow the classifier to focus on the morphological and chromatic attributes of lesions rather than on spurious variability introduced by acquisition conditions. This phase thus plays a central role in enabling the model to learn discriminative features reliably across the ten diagnostic categories present in the dataset.

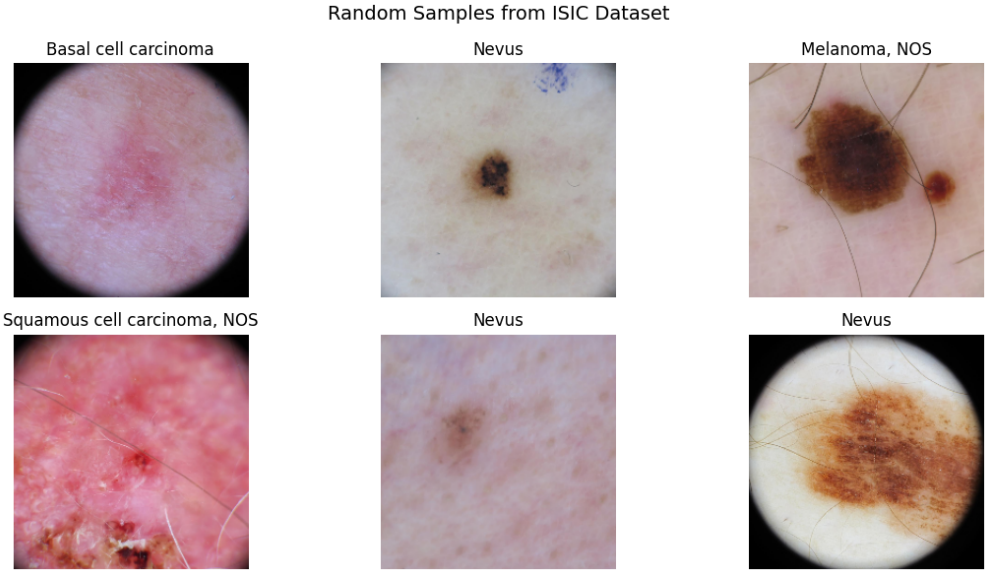


Fig. 1. Some randomly chosen preprocessed images from dataset

### 3.4 Phase 2: Model Architecture Selection and Training Setup

The core of the classification system is a deep convolutional neural network designed to learn discriminative visual features from dermoscopic images. In this study, an EfficientNet backbone was employed as the primary feature extractor due to its strong performance–efficiency trade-off and its demonstrated suitability for fine-grained visual recognition tasks. EfficientNet's compound scaling strategy allows width, depth and resolution to be balanced in a principled manner, enabling the network to model subtle morphological distinctions between lesion categories without incurring unnecessary computational overhead. The model was initialized with ImageNet-pretrained weights, leveraging transfer learning to accelerate convergence and to counteract the limited sample size of rare diagnostic classes.

During the initial training phase, the majority of the EfficientNet layers were frozen, allowing only the newly added classification head to be optimized. This head consisted of a global average pooling layer followed by one or more dense layers and a final softmax output layer corresponding to the ten diagnostic categories. Freezing the backbone stabilizes early training dynamics by preventing catastrophic weight updates to pretrained convolutional layers before the classification head has adapted to the dermoscopic domain. After this warm-up stage, selective fine-tuning was performed by unfreezing deeper convolutional blocks, enabling the network to adjust high-level representations to lesion-specific texture and pigmentation patterns. This two-step

approach balances generalization and specialization, ensuring that the pretrained visual priors are preserved while allowing the network to adapt to domain-specific cues.

Training was conducted using the Adam optimizer with categorical cross-entropy loss, a combination suitable for multi-class classification and effective in handling imbalanced datasets when coupled with class-aware sampling or augmentation strategies. The learning rate schedule included an initial low learning rate to protect pretrained layers during fine-tuning, followed by gradual decay to refine the model's decision boundaries. Additional mechanisms such as early stopping and model checkpointing were incorporated to prevent overfitting, monitor validation performance and retain the best-performing weights. Batches were constructed using the preprocessed and augmented TensorFlow pipeline, ensuring efficient GPU utilization and consistent input quality throughout training.

Overall, this architecture and training strategy provide a strong foundation for multi-class dermoscopic classification by combining the representational power of transfer learning with domain-specific fine-tuning. The next phase evaluates the predictive behavior of the trained model, examining both aggregate classification metrics and class-wise diagnostic tendencies to assess clinical viability.

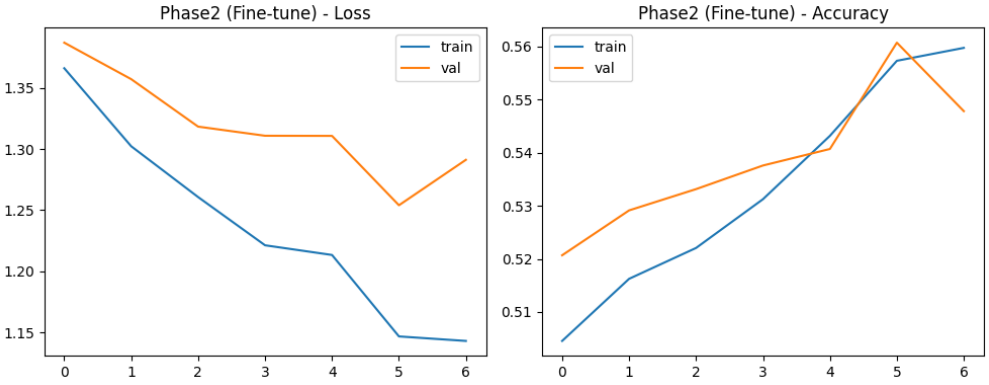


Fig. 2. Training and validation learning curves during Phase 2 fine-tuning

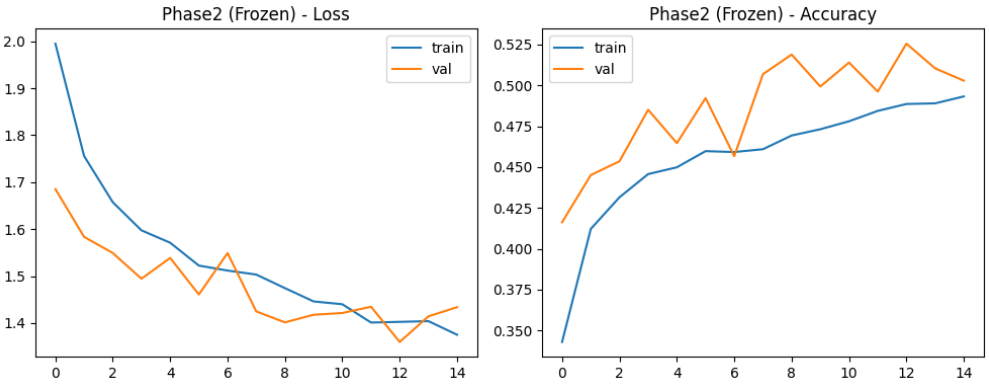


Fig. 3. Training and validation loss and accuracy trends during Phase 2 with frozen EfficientNet backbone

### 3.5 Phase 3: Model Evaluation Framework

The evaluation of the trained classifier was carried out using a multi-metric framework designed to assess both overall predictive performance and class-specific diagnostic behavior. Since dermoscopic images exhibit substantial visual heterogeneity and the dataset contains pronounced class imbalance, reliance on a single metric such as accuracy would provide an incomplete picture of model reliability. Therefore, a suite of complementary measures was adopted, including precision, recall, F1-score and the confusion matrix, each of which highlights different aspects of the model's discriminative capacity.

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Basal cell carcinoma | 0.652062 | 0.459165 | 0.538871 | 551.000000 |
| Dermatofibroma | 0.272727 | 0.720000 | 0.395604 | 25.000000 |
| Melanoma metastasis | 0.268398 | 0.652632 | 0.380368 | 95.000000 |
| Melanoma, NOS | 0.695652 | 0.505824 | 0.585742 | 601.000000 |
| Nevus | 0.785915 | 0.658796 | 0.716763 | 847.000000 |
| Scar | 0.126984 | 0.510638 | 0.203390 | 47.000000 |
| Seborrheic keratosis | 0.339623 | 0.378947 | 0.358209 | 190.000000 |
| Solar lentigo | 0.237288 | 0.651163 | 0.347826 | 43.000000 |
| Solar or actinic keratosis | 0.405714 | 0.435583 | 0.420118 | 163.000000 |
| Squamous cell carcinoma, NOS | 0.291667 | 0.416667 | 0.343137 | 84.000000 |
| accuracy | 0.538549 | 0.538549 | 0.538549 | 0.538549 |
| macro avg | 0.407603 | 0.538941 | 0.429003 | 2646.000000 |
| weighted avg | 0.622332 | 0.538549 | 0.563851 | 2646.000000 |

Fig. 4. Class-wise precision, recall and F1-scores for the trained model

Accuracy was used as a coarse indicator of general performance across the ten diagnostic categories. However, because frequent classes such as nevus dominate the dataset, precision and recall were computed for each class to capture asymmetries in false-positive and false-negative rates. Precision evaluates the proportion of predicted samples that are truly positive for each class, which is particularly important for categories prone to over-prediction due to visual similarity. Recall, in contrast, measures the ability of the classifier to correctly identify all relevant samples from a given lesion type. The F1-score provides a balanced harmonic mean of these two quantities, mitigating distortions caused by class imbalance and making it suitable for comparing performance across rare and common lesions alike.

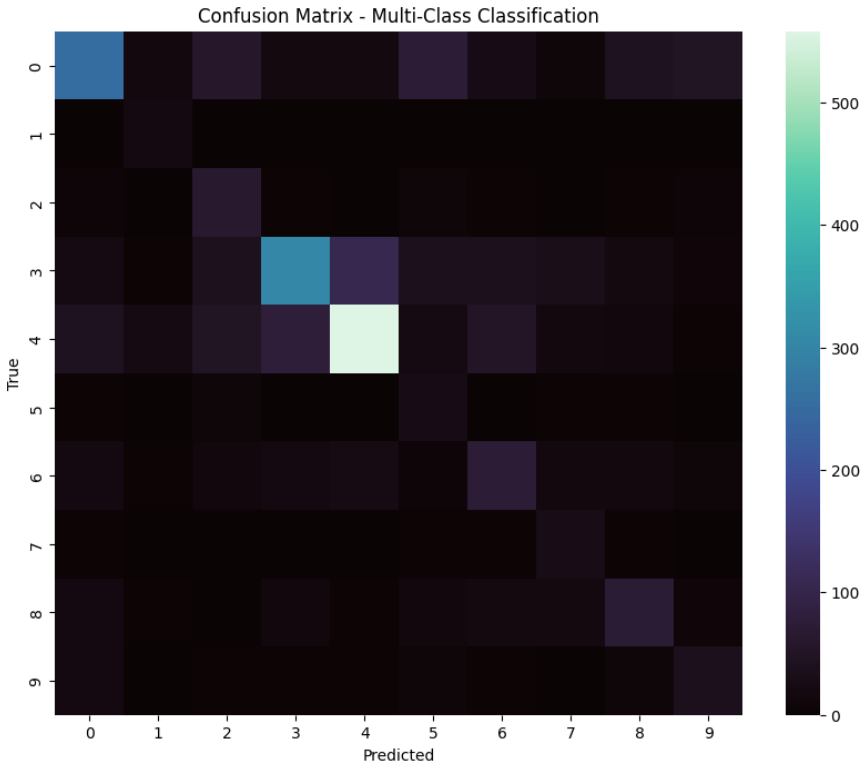


Fig. 5. Confusion matrix for the ten-class skin lesion classifier

A confusion matrix was generated to visualize the distribution of model errors across diagnostic categories. This matrix functions as a detailed error map that reveals systematic misclassifications, such as melanoma being confused with atypical nevi or seborrheic

keratoses being mistaken for solar lentigines. These patterns carry clinical significance, as they may reflect underlying feature similarities between lesion types or model biases introduced by sample scarcity. The confusion matrix also serves as a foundation for the subsequent explainability analysis, since misclassified instances were later examined using Grad-CAM and Grad-CAM++ to determine whether errors arose from ambiguous morphology, irrelevant image regions or non-dermatological artefacts.

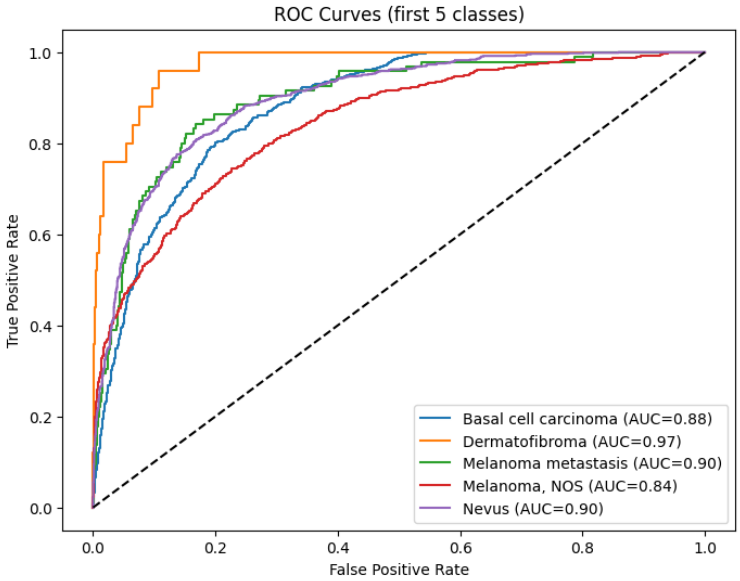


Fig. 6. ROC curves for five representative lesion classes

During training, evaluation was performed at the end of each epoch using the held-out validation set to monitor convergence behavior and detect early signs of overfitting. Model checkpointing ensured that the best-performing weights were retained based on validation accuracy, while early stopping prevented unnecessary training epochs. Final evaluation was conducted on the independent test set to provide an unbiased estimate of the classifier's performance after full training and fine-tuning. The combined use of metric-based evaluation and error-structure analysis offers a comprehensive understanding of the model's strengths and limitations, setting the stage for integrating explainability methods that can further contextualize the classifier's decision-making patterns.

### 3.6 Phase 4: Explainability Pipeline (Grad-CAM and Grad-CAM++)

Explainability plays a central role in evaluating whether a deep learning model's predictions align with clinically meaningful visual cues. In this phase, post-hoc interpretability methods were applied to reveal the internal reasoning of the EfficientNet classifier and to examine whether the features driving predictions correspond to dermatologically relevant structures. The analysis employed Grad-CAM and Grad-CAM++, two widely used class-discriminative visualization techniques that produce heatmaps highlighting the image regions most influential for the predicted class.

Grad-CAM operates by tracing the gradient of the target class score back to the final convolutional feature maps of the network. These gradients quantify the contribution of each spatial location to the model's decision, enabling the construction of a coarse localization map reflecting class-specific attention. For this study, gradients were extracted from the final convolutional block of EfficientNet, as this layer captures high-level semantic features such as pigment asymmetry, border irregularity and global texture patterns that dermatologists use for diagnostic assessment. The resulting heatmaps were normalized and superimposed onto the original dermoscopic images to create intuitive visual explanations that clinicians and researchers can interpret without specialized knowledge of neural network internals.

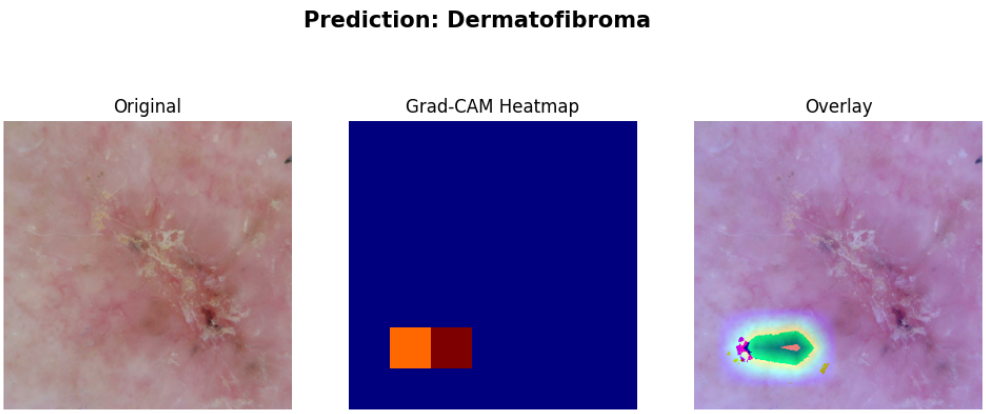


Fig. 7. Grad-CAM explanation for a misclassified dermatofibroma sample

While Grad-CAM provides valuable insights, its reliance on first-order gradients can limit its ability to distinguish fine-grained feature contributions in cases with multiple discriminative regions. To address this limitation, Grad-CAM++ was additionally employed. Grad-CAM++ incorporates second-order derivatives to better handle overlapping features and to distribute attention across multiple relevant areas rather than focusing primarily on a single region of interest. This leads to sharper and more detailed localization maps, particularly beneficial in complex lesions such as melanoma, where diagnostic cues may appear across irregular pigment networks or peripheral streaks. By comparing Grad-CAM and Grad-CAM++ outputs, the analysis could differentiate between coarse and fine-grained interpretive signals, improving both the fidelity and stability of the explanations.

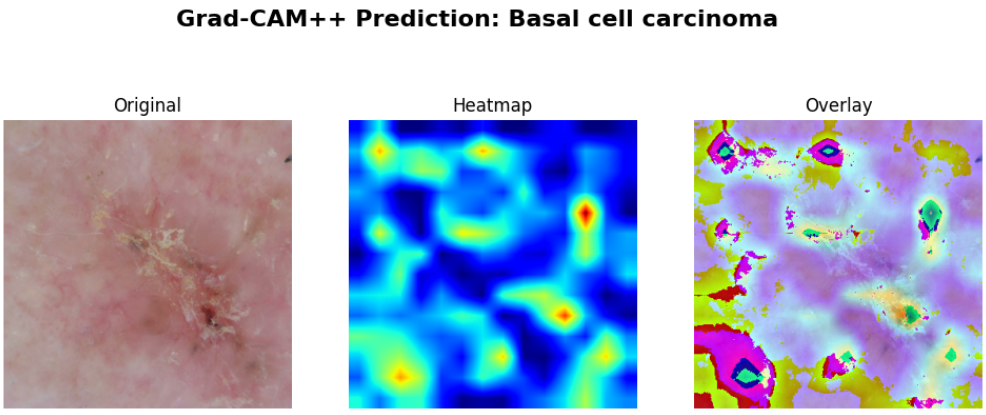


Fig. 8. Grad-CAM++ explanation for a correctly classified basal cell carcinoma

The explainability pipeline was applied to correctly classified images, borderline cases and misclassifications to contextualize the model's behavior. For accurate predictions, heatmaps often highlighted clinically appropriate regions such as the lesion core or pigment-dense structures, suggesting that the classifier relied on legitimate dermatological patterns. In contrast, misclassified samples frequently revealed attention shifts toward irrelevant artefacts, image borders or background noise, offering insight into failure modes that cannot be detected through performance metrics alone. These interpretive observations provided a deeper understanding of error sources, such as morphological ambiguity between lesion subtypes or biases introduced by class imbalance, and directly informed the error analysis conducted in Phase 5.

Through the combined use of Grad-CAM and Grad-CAM++, this phase establishes a transparent interpretability layer atop the EfficientNet classifier. The resulting visualizations do not merely justify individual predictions but also contribute to assessing the model's overall reliability, identifying systematic weaknesses and guiding future improvements in both dataset curation and model design.

### 3.7 Phase 5: Error Analysis

A comprehensive error analysis was conducted to better understand the strengths and limitations of the trained classifier and to assess how misclassifications arise across lesion categories. While standard metrics such as accuracy and F1-score provide broad indicators of performance, they do not reveal whether the model's errors stem from clinically plausible similarities between lesions, dataset imbalances or attention to irrelevant features. To address this, multiple complementary tools were used, including class-wise ROC curves, the confusion matrix and visual explanations generated through Grad-CAM and Grad-CAM++. Together, these analyses provide a multi-layered perspective on predictive behavior and failure modes.

Class-wise ROC curves offer insight into how well the model separates each lesion type from the remaining classes. The results demonstrate substantial variability in discriminative performance: dermatofibroma achieves a very high AUC, while melanoma subtypes show more modest separability. This pattern is consistent with the underlying dataset distribution, where some classes exhibit distinctive structural features while others overlap visually with multiple categories. ROC curves therefore contextualize performance differences as a function of intrinsic class similarity rather than model instability.

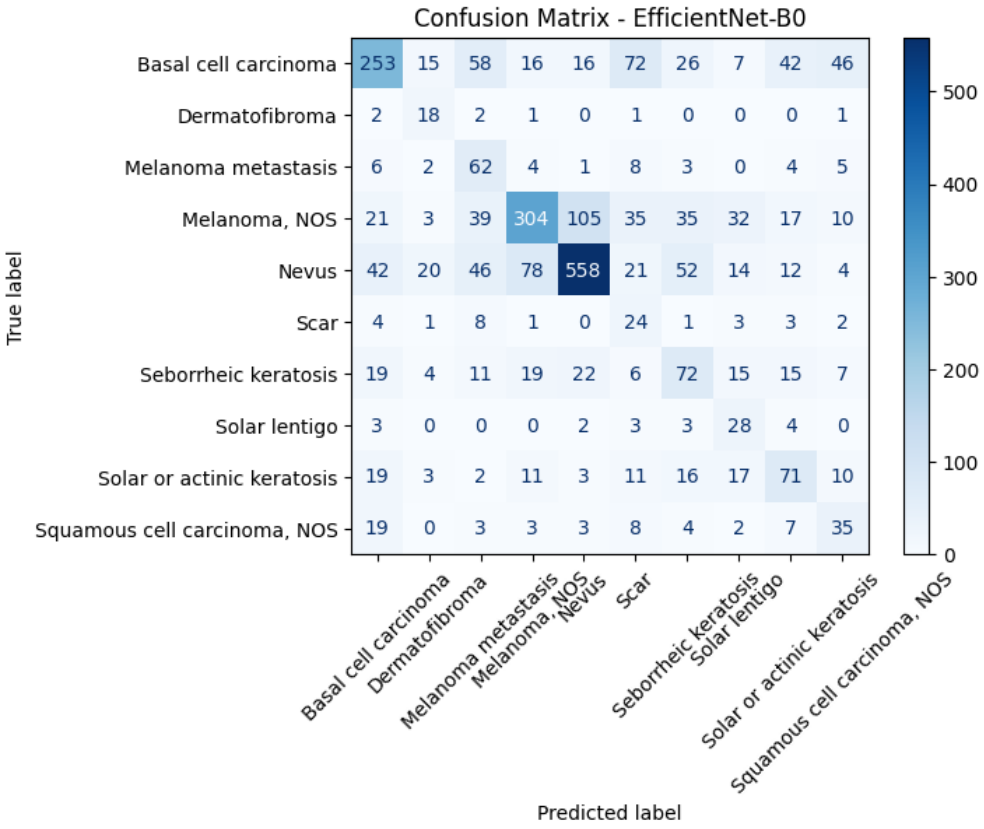


Fig. 9. Confusion matrix showing class-wise error patterns of EfficientNet-B0 classifier

The confusion matrix further exposes systematic error patterns at the class level. High concentration along the diagonal reflects correct predictions, but off-diagonal clusters reveal clinically meaningful misclassifications. For example, melanoma cases are often confused with atypical nevi, and solar lentigo may be misinterpreted as seborrheic keratosis, mirroring the real-world diagnostic difficulty of differentiating these lesions. Conversely, rare categories such as scar and melanocytic metastasis show inconsistent predictions due to limited sample representation. These patterns indicate that the classifier's behavior reflects both dermatological ambiguity and dataset imbalance, necessitating careful interpretation of classification metrics.

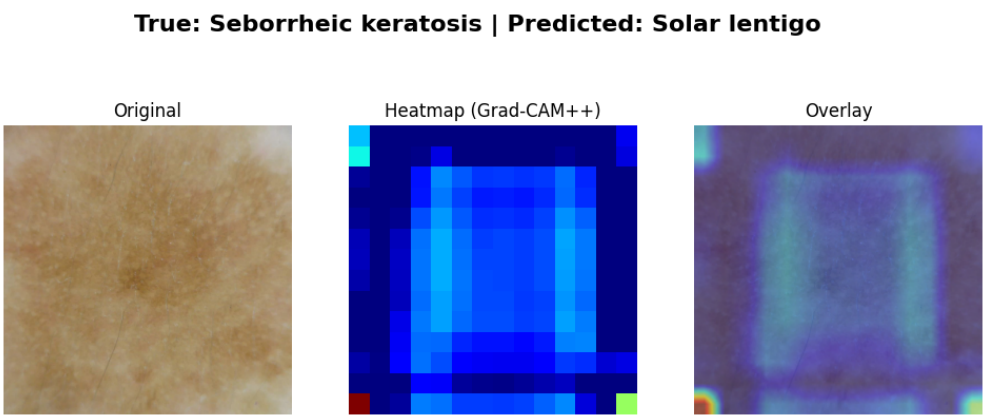


Fig. 10. Grad-CAM++ explanation for a misclassified Seborrheic keratosis sample (predicted as Solar lentigo)

The explainability methods were then applied to individual test cases to inspect the visual basis of both correct and incorrect predictions. In correctly classified samples, Grad-CAM and Grad-CAM++ heatmaps typically focused on diagnostically relevant regions within the lesion, such as pigment networks, asymmetries or textural irregularities. This alignment suggests that the model learned meaningful clinical cues rather than relying on incidental artefacts. In contrast, misclassified samples frequently exhibited attention drift toward irrelevant regions, including image borders, hairs, reflections or isolated background structures. These findings demonstrate that some errors arise not from morphological ambiguity but from the model anchoring its predictions on spurious patterns unrelated to the lesion.

The combination of quantitative and qualitative error analysis provides crucial insight into the model's reliability. ROC and confusion matrix findings reveal where the classifier struggles at a class-wide level, while explainability outputs clarify why individual predictions succeed or fail. By identifying cases where the model relies on artefactual features or fails to attend to the lesion core, this phase highlights important avenues for improvement, such as enhanced data cleaning, balanced sampling strategies and more robust attention mechanisms. Ultimately, the error analysis affirms that explainability is not only a tool for interpreting predictions but also a diagnostic instrument for refining model design and dataset quality.

## 4 Results

The classifier demonstrated moderate performance on the ten-class dermoscopic dataset, achieving an overall accuracy of 53.8% on the held-out test set. Performance varied substantially across categories, reflecting both intrinsic visual similarity between certain lesions and dataset imbalance. Nevus and melanoma subtypes achieved the highest F1-scores among common classes, while rare categories such as scar and dermatofibroma showed reduced precision due to limited sample representation. Class-wise ROC curves revealed generally strong discriminative ability, with AUC values ranging from 0.84 to 0.97 for the most frequent five classes.

The confusion matrix exposed consistent misclassification patterns, particularly between visually overlapping lesions such as seborrheic keratosis and solar lentigo, or melanoma and atypical nevi. These trends align with known diagnostic challenges in dermatology. Explainability visualizations provided further insight into model behavior. In correctly classified images, Grad-CAM and Grad-CAM++ heatmaps highlighted clinically meaningful structures within the lesion. In contrast, misclassified samples often showed attention drifting toward non-diagnostic regions, indicating that some errors were driven by artefacts rather than lesion morphology.

## 5 Discussion

The results suggest that convolutional models can extract meaningful features from dermoscopic images but remain sensitive to dataset limitations and class imbalance. Strong AUC values for several lesion types indicate that the network learned discriminative patterns, yet the modest overall accuracy reflects the difficulty of fine-grained dermatological classification without extensive data augmentation or balanced sampling strategies. Error patterns observed in the confusion matrix show that the classifier struggles most with classes that also pose challenges in clinical practice, which supports the realism of the task.

The explainability analysis proved essential for understanding model failures. Grad-CAM and Grad-CAM++ revealed that incorrect predictions frequently stemmed from reliance on artefacts such as hairs, borders or illumination irregularities. These insights highlight the importance of incorporating interpretability into medical AI systems, not only to justify predictions but also to guide dataset refinement and model redesign. The findings also suggest that integrating lesion segmentation, additional augmentation techniques or attention-based architectures may reduce attention drift and improve robustness.

## 6 Conclusion

This study developed a deep learning model for multi-class skin lesion classification and integrated explainability tools to examine the decision-making process. The classifier achieved competitive class-wise discrimination on several lesion categories, and the interpretability analysis revealed how the model used clinically meaningful structures for correct predictions while exposing failure modes in misclassified cases. Although overall accuracy remains limited by dataset imbalance and the complexity of dermoscopic images, the combined use of performance metrics and visual explanations provides a transparent framework for evaluating and improving dermatological AI systems.

Future work may incorporate segmentation-enhanced training, balanced sampling strategies and larger datasets, along with clinician-in-the-loop evaluation to further assess the utility of explanation maps in real diagnostic settings.

## References

[1] O. S. Ekundayo and A. E. Ezugwu, "Deep learning: Historical overview from inception to actualization, models, applications and future trends," in *Applied Soft Computing*, 2025. [Online]. Available: https://doi.org/10.1016/j.asoc.2025.113378

[2] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," vol. 8, no. 53, 2021, pp. 1–74. [Online]. Available: https://doi.org/10.1186/s40537-021-00444-8

[3] Y. Li, H. Zhang, X. Xue, Y. Jiang, and Q. Shen, "Deep learning for remote sensing image classification: A survey," in *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 6, 2018, p. e1264. [Online]. Available: https://doi.org/10.1002/widm.1264

[4] P. E. D. Love, W. Fang, J. Matthews, S. Porter, H. Luo, and L. Ding, "Explainable artificial intelligence (xai): Precepts, models, and opportunities for research in construction," in *Advanced Engineering Informatics*, vol. 57, 2023, p. 102024. [Online]. Available: https://doi.org/10.1016/j.aei.2023.102024

[5] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," vol. 6, no. 6, 2020, p. 52. [Online]. Available: https://doi.org/10.3390/jimaging6060052

[6] S. P. Dharshini, K. R. Kumar, S. Venkatesh, K. Narasimhan, and K. Adalarasu, "An overview of interpretability techniques for explainable artificial intelligence (xai) in deep learning-based medical image analysis," in *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*. IEEE, 2023, pp. 175–182. [Online]. Available: https://doi.org/10.1109/ICACCS57279.2023.10113001

[7] S. Solanki, U. P. Singh, and S. S. Chouhan, "Brain tumor classification using ML and DL approaches," in *2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*. IEEE, 2023, pp. 204–208. [Online]. Available: https://doi.org/10.1109/ICCCMLA58983.2023.10346854

[8] K. M. Hosny, W. Said, M. Elmezain, and M. A. Kassem, "Explainable deep inherent learning for multi-classes skin lesion classification," vol. 159, 2024, p. 111624. [Online]. Available: https://doi.org/10.1016/j.asoc.2024.111624

[9] N. Veeramani, R. S. Sherine, S. Prabha, S. Srinidhi, and P. Jayaraman, "Nextgen lung disease diagnosis with explainable artificial intelligence," in *Scientific Reports*, 2025. [Online]. Available: https://doi.org/10.1038/s41598-025-07603-4

[10] M. Arshad, M. A. Khan, N. A. Almujally, A. Alasiry, M. Marzougui, and Y. Nam, "Multiclass skin lesion classification and localization from dermoscopic images using a novel network-level fused deep architecture and explainable artificial intelligence," in *BMC Medical Informatics and Decision Making*, vol. 25, 2025, p. 215. [Online]. Available: https://doi.org/10.1186/s12911-025-03051-2

[11] N. Ahmad, J. H. Shah, M. A. Khan, J. Baili, G. J. Ansari, U. Tariq, Y. J. Kim, and J.-H. Cha, "A novel framework of multiclass skin lesion recognition from dermoscopic images using deep learning and explainable AI," in *Frontiers in Oncology*, vol. 13, 2023, p. 1151257. [Online]. Available: https://doi.org/10.3389/fonc.2023.1151257

[12] I. Jahan, A. H. Efat, S. M. M. Hasan, M. F. Faruk, A. Y. Srizon, M. R. Hossain, and M. A. Mamun, "An explainable deep learning framework for multi-class skin lesion classification while resolving class imbalance," in *2024 IEEE International Conference on Power, Electrical, Electronics and Industrial Applications (PEEIACON)*, 2024, pp. 1–6. [Online]. Available: https://doi.org/10.1109/PEEIACON63629.2024.10800046

[13] N. Nigar, M. Umar, M. K. Shahzad, S. Islam, and D. Abalo, "A deep learning approach based on explainable artificial intelligence for skin lesion classification," in *IEEE Access*, vol. 10, 2022, pp. 113 715–113 725. [Online]. Available: https://doi.org/10.1109/ACCESS.2022.3217217