Interpretable AI for Skin Screening: MultiExCam for Melanoma, Basal Cell Carcinoma and Benign Nevi Classification

Maira Aracne[1], Luciano Caroprese[2], Tommaso Ruga[3,*], Eugenio Vocaturo[4] and Ester Zumpano[3]

[1]University Leonardo da Vinci, Piazza San Rocco, 2, Torrevecchia Teatina, Italy
[2]Engineering and Geology Department, University G. d'Annunzio of Chieti-Pescara, Viale Pindaro 42, Pescara, Italy
[3]DIMES Department, University of Calabria, Via Ponte Pietro Bucci, Rende (CS), Italy
[4]CNR-NANOTEC, National Research Council, Rende (CS), Italy

Abstract

Accurate differentiation among melanoma, basal cell carcinoma (BCC), and benign nevi remains challenging in clinical dermatology, despite their distinct prognostic implications. While recent advances in artificial intelligence have improved binary skin lesion classification, multiclass scenarios reflecting real-world diagnostic complexity remain still challenging. This study extends MultiExCam, a hybrid deep learning and machine learning framework, to address multiclass classification with enhanced explainability for clinical decision support. By incorporating basal cell carcinoma as a third diagnostic class, we demonstrate how hybrid integration of deep and machine learning can effectively address real-world multiclass skin lesion scenarios. Explainability analysis via Grad-CAM and SHAP enables identification of clinically meaningful features distinguishing melanoma (asymmetry, irregular borders, color variation), BCC (vascular patterns, pearly appearance, ulceration), and nevi (symmetry, homogeneity), providing interpretable decision rationales aligned with dermatological diagnostic criteria. The framework's ability to differentiate among three distinct lesion types with transparent decision-making processes addresses critical requirements for clinical adoption, offering a promising foundation for AI-assisted dermatology and screening applications in resource-limited settings.

Keywords

Skin lesion classification, Medical Image Analysis, Explainable AI, Transfer learning, Ensemble learning

1. Introduction

Cutaneous melanoma is among the deadliest forms of skin cancer, with the World Health Organization [1] reporting over 320,000 new cases and approximately 58,000 deaths annually, a number that is expected to grow. In Europe, the incidence has steadily increased, reaching 10-25 new cases per 100,000 inhabitants by 2022 [2]. Early detection remains crucial, as timely diagnosis significantly improves treatment outcomes and patient survival. Accurately classifying skin lesions, however, remains a complex task. While benign lesions such as *nevi* typically exhibit symmetry and regular borders, malignant lesions like *melanoma* often show asymmetry and irregular edges. Although distinguishing melanoma from nevi is clinically critical, other skin lesion types also warrant accurate identification. *Basal cell carcinoma* (BCC), the most common form of skin cancer, originates from basal cells in the epidermis. While it rarely metastasizes, BCC can cause extensive local tissue damage if left untreated [3]. A comparison of the three lesion types is illustrated in Figure 1.

Motivated by these considerations, this study extends the traditional binary classification task of melanoma versus nevus to a multi-class setting that includes BCC alongside other lesion types. Specifically, we build upon our existing framework *MultiExCAM* [4], a hybrid and explainable architecture for skin lesion classification, originally developed for melanoma/nevi discrimination, enhancing it to support the identification of three different skin lesions types simultaneously. This extension aims

Published in the Proceedings of the Workshops of the EDBT/ICDT 2026 Joint Conference (March 24-27, 2026), Tampere, Finland
*Corresponding author.
m.aracne@unidav.it (M. Aracne); luciano.caroprese@unich.it (L. Caroprese); tommaso.ruga@dimes.unical.it (T. Ruga); eugenio.vocaturo@cnr.it (E. Vocaturo); e.zumpano@dimes.unical.it (E. Zumpano)

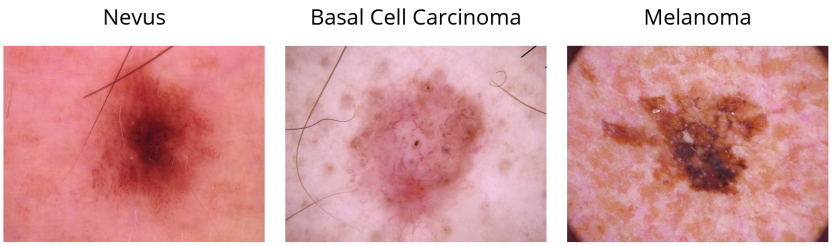

Figure 1: The visual differences between common nevi, basal cell carcinoma, and melanomas.

to provide a more comprehensive automated diagnostic tool, improving early detection and assisting clinicians in managing a wider spectrum of skin cancers.

In addition, this study is motivated by a set of simple yet critical research questions: although the approaches proposed in the literature report promising performance, do they provide an adequate level of result explainability? Are convolutional neural networks still a valid architectural choice for building classifiers in light of the recent emergence of Vision Transformers? Consequently, does a model such as MultiExCam remain effective when moving beyond a binary classification setting to a multi-class scenario?

The paper is structured as follows. In the next section, a comprehensive review of the state of the art is presented, aiming to provide an overview of current techniques and establish a baseline for comparison. Section 3 introduces the MultiExCam architecture, detailing its individual components and design choices. In Section 4, the experimental results are reported and critically analyzed. Finally, Section 5 concludes the study and discusses potential directions for future research.

2. Related Works

Several studies have proposed architectures capable of distinguishing among multiple types of skin lesions. In [5], the authors proposed SKINC-NET, a lightweight deep learning architecture designed for the accurate classification of multiple skin cancer types while maintaining low computational complexity. The model is capable of classifying seven different skin lesion categories from dermoscopic images and is built using dense layers, batch normalization, and LeakyReLU activation functions. To address the class imbalance issue, data augmentation techniques were applied and the model was evaluated on the HAM10000 dataset. The proposed approach was compared with several transfer learning models, including ResNet50, VGG16, MobileNetV2, and EfficientNetB0. Furthermore, ablation studies were conducted to determine optimal hyperparameter settings. Experimental results demonstrated that SKINC-NET achieved superior performance, reaching an overall accuracy of 98.54% with a significantly reduced number of trainable parameters and computational cost. Also in [6], the dataset HAM10000 was used to evaluate three pre-trained CNN architectures (CNN, VGG16, VGG19) and three hybrid models combining these networks (CNN+VGG16, CNN+VGG19, VGG16+VGG19). The proposed multilayered hybrid model, designed with carefully varied filter sizes while maintaining a reduced number of parameters, achieved the highest performance, reaching an accuracy of 91.63%. Model effectiveness was also assessed using precision, recall, and F1-score metrics, demonstrating its capability to support automated skin lesion diagnosis in a multi-class setting. In [7], the authors proposed a hybrid deep learning architecture for skin cancer classification that integrates ConvNeXtV2 blocks with separable self-attention mechanisms to enhance feature extraction and classification performance. The model leverages ConvNeXtV2 blocks in the early stages to capture fine-grained local features and subtle visual patterns, which are crucial for discriminating between visually similar skin lesions. In later stages, separable self-attention is employed to focus on diagnostically relevant regions with reducing computational overhead. The proposed approach was trained and evaluated on the ISIC 2019 dataset, encompassing eight distinct skin lesion categories, and was further strengthened through the use of data augmentation and transfer learning techniques. The proposed model achieved an accuracy of 93.48%, a precision of 93.24%, a recall

of 90.70%, and an F1-score of 91.82%, outperforming over ten CNN-based and more than ten Vision Transformer-based models evaluated under similar conditions. Again, in [8], authors used ISIC2019 dataset to test different hybrid architecture, proposing SkinDWNet, a deep learning architecture for multiclass skin cancer classification that combines depth-wise dilated convolutions (DDCs) with feature reuse residual blocks (FRBs) to effectively extract relevant features from dermoscopic images. To enhance classification performance, the model is coupled with Gradient Boosting, which leverages the feature maps generated by SkinDWNet. The study also employs preprocessing techniques to remove artifacts such as hair and air bubbles, as well as SMOTE Tomek to address class imbalance in the ISIC 2019 dataset. SkinDWNet+GB was evaluated against six baseline and state-of-the-art models, achieving a maximum accuracy of 97.04%, and statistical analyses including ANOVA and McNemar's test confirmed its performance. In addition, a Grad-CAM analysis was proposed in the work. In [9], the authors proposed a deep learning framework for multi-class classification of dermoscopic skin lesions using the HAM10000 dataset. Several state-of-the-art CNN architectures, including DenseNet201, InceptionResNetV2, and Xception, were evaluated under both frozen and fully fine-tuned settings, and Vision Transformer (ViT) models were also assessed for their potential in skin lesion analysis. To further improve performance, ensemble learning strategies such as hard voting, soft voting, and weighted soft voting were applied. Experimental results showed that fully fine-tuned models outperformed frozen configurations, with InceptionResNetV2 achieving the best individual performance (accuracy 88%, F1-score 0.77), while the weighted soft voting ensemble reached the highest overall results (accuracy 89%, F1-score 0.80). Examining the state of the art allows us to address two of the research questions posed earlier. CNNs remain a valid approach for image analysis, although in some solutions they are complemented or replaced by more recent architectures, such as ViTs. A persistent concern, however, is explainability: only a small fraction of existing studies employ explainable techniques to justify their predictions. To address this limitation, MultiExCam, even in its initial version, is designed as a framework that is inherently explainable and capable of providing reliable predictions regarding the nature of skin lesions.

3. Methodology

The MultiExCam framework, originally proposed in [10, 4] is designed to integrate deep learning (DL) and traditional machine learning (ML) techniques to achieve high classification performance while maintaining high level of results interpretability. The framework is structured in four main phases. First of all, raw dermoscopic images are standardized for analysis through operations such as resizing and artifact removal (e.g., occlusions caused by body hair), ensuring clean inputs for subsequent stages. The images are then fed into the DL block where a pre-trained CNN, a ResNet50, fine-tuned via transfer learning, performs two parallel tasks: i) a coarse initial classification providing a baseline prediction of lesion type, and (ii) deep feature extraction, where high-level representations are obtained from the penultimate fully connected layer. These extracted features serve as inputs for the ML classifiers in the next phase. In the ML block, four complementary ML classifiers are applied to a hybrid feature set that combines CNN-extracted deep features with handcrafted descriptors, such as statistical and colorimetric features. The diversity of classifiers, designed to cover both linear and nonlinear decision boundaries, enhances robustness and reliability of intermediate predictions. The final block is constituted by a feedforward neural network (FFNN) meta-classifier aggregates the outputs of individual classifiers and the hybrid features, acting as a specialized classifier, similar to a medical expert in a clinical protocol, which receives information from the other models and produces the final classification by weighing their contributions. Its multi-branch design, equipped with an attention mechanism, learns optimal fusion strategies, weighing the contribution of each classifier to generate refined, context-aware final predictions. To ensure interpretability, MultiExCam incorporates two explainable AI techniques: Grad-CAM for image-level visualization and SHAP for feature-level transparency. This combination provides clinicians with insight into the decision-making process, enhancing trust and supporting accurate diagnostic decisions. A framework overview is proposed in Figure 2.

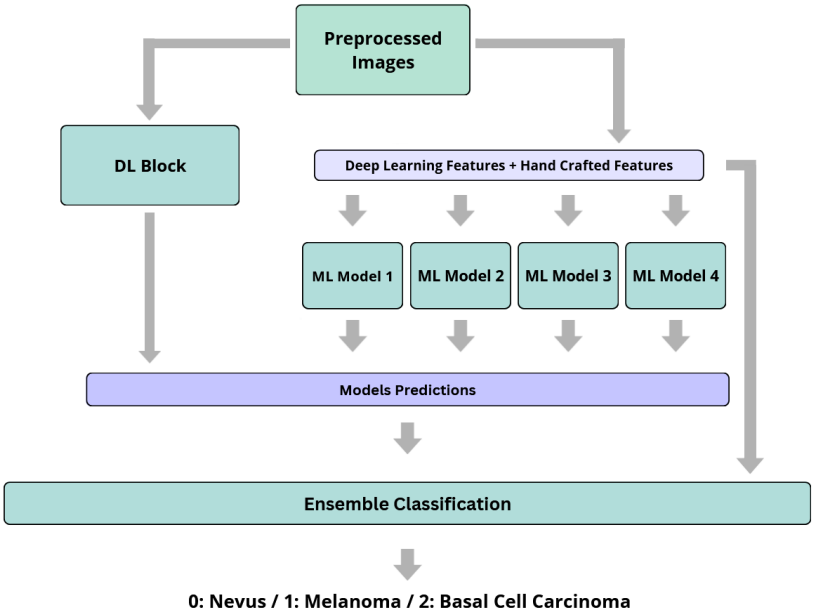

Figure 2: *The MultiExCam architecture is characterized by a hybrid design that employs three distinct classification stages: DL, ML, and ensemble.*

3.1. Dataset

The original MultiExCam framework was validated using three different datasets: HAM10000, the Skin Cancer Detection dataset (Kaggle), and MED-NODE. Since only the first two datasets contain the BCC class, and only the first provides a sufficiently large number of BCC images, the experimental evaluation in this study was conducted solely on the HAM10000 dataset. The *HAM10000* dataset [11], released as part of the ISIC 2018 Challenge [12], consists of 10,000 RGB dermoscopic images with a resolution of 600×450 pixels. For this study, we focused on three skin lesion classes: *melanoma* (1,134 images), *basal cell carcinoma* (BCC, 529 images), and *common nevus* (6,705 images). The official training and validation sets were merged, randomly shuffled, and then partitioned into new training, validation, and test sets using a 70-15-15 split. The dataset is highly imbalanced, with nevi representing the majority class. To address this imbalance, a two-pronged strategy was applied to both the training and validation sets. Specifically, the number of nevi images was reduced to match that of melanomas, while the number of BCC images was increased via geometric transformations. This approach brought all three classes to an equal number of samples, while the test set remained untouched to provide an unbiased evaluation of model performance.

4. Results

We conducted a series of experiments on the Google Colab platform with a T4 GPU runtime and high RAM to assess the efficiency and accuracy of MultiExCam, following the same experimental setup as the original paper. Details of these experiments will be discussed in the following section.

Table 1 reports the classification performance of the proposed ensemble model on the test set. The results are presented per class, along with the weighted average across all classes. These metrics reflect the effectiveness of the ensemble in combining the predictions of individual models and provide a quantitative assessment of its predictive performance on unseen data.

The proposed model achieved a weighted average precision of 86%, recall of 85%, and F1-score of 86% across the three classes. In comparison, the original binary version of the framework reported weighted average metrics of 92% for precision, recall, and F1-score. While the introduction of a third class leads to a moderate decrease in performance, the decline is not substantial, indicating that the model maintains

Table 1
Classification metrics per class.

Class	Precision	Recall	F1-score
Nevus (0)	91%	87%	89%
Melanoma (1)	84%	85%	84%
BCC (2)	78%	84%	81%
Weighted avg	86%	85%	86%

robust predictive capabilities even under a more challenging multiclass scenario. To further analyze the model's performance, we present two confusion matrices on the test set. The first corresponds to the ResNet50 model alone (Fig. 3), while the second shows the predictions after the ensemble block (Fig. 4).

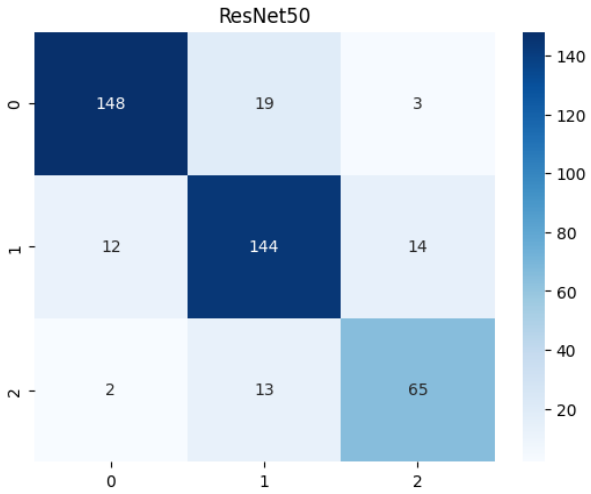

Figure 3: Confusion matrix obtained on the test set after classification with the DL block.

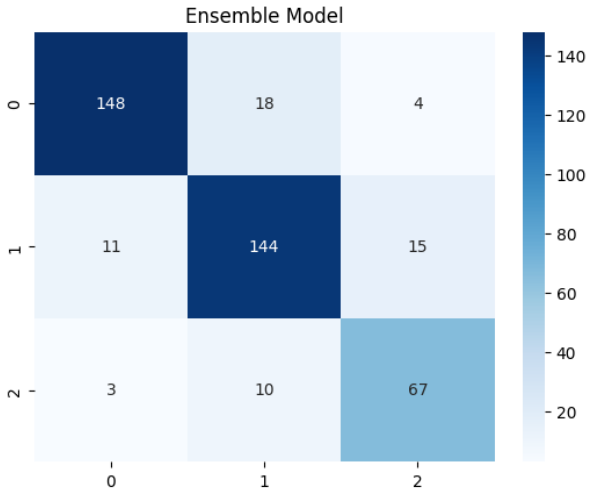

Figure 4: Confusion matrix obtained on the test set after ensemble classification process.

These matrices provide a detailed breakdown of correct and incorrect classifications for each class, highlighting the specific patterns of misclassification. This comparison allows for a clearer assessment of the improvements introduced by the ensemble approach and complements the summary metrics reported in Table 1. The confusion matrix shown in Figure 4 demonstrates improved recognition of the newly introduced class (BCC, class 2), correctly classifying 67 out of 80 samples, which represents

a substantial improvement compared to individual models. For the original classes (Nevus, class 0, and Melanoma, class 1), the ensemble maintains a performance comparable to the single ResNet50 model, correctly classifying 148 out of 170 Nevus samples and 144 out of 170 Melanoma samples. These results indicate that the ensemble effectively balances the recognition of the new class while preserving the accuracy on previously established classes. From an explainability perspective, the Grad-CAM technique [13], already employed in the original binary framework, was applied to the output of DL module. The resulting activation maps reveal that the model consistently focuses on the lesion area correctly attributing importance to the relevant features, as could be noted looking at the warm area present in the heatmaps. Notably, this behavior is also observed for the newly introduced BCC class, as illustrated in Figure 5.

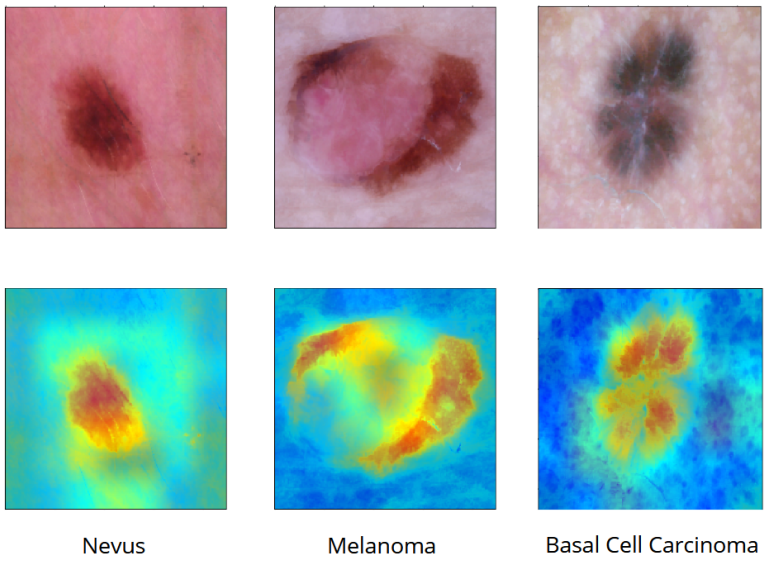

Figure 5: Grad-CAM analysis revealed consistent and rigorous behavior across all classes

Beyond confirming the model's attention to clinically meaningful regions, the explainability analysis also allowed us to identify two recurring sources of difficulty. First, images in which the lesion area is not clearly visible (Figure 6 (a)), making it inherently harder for the model to extract discriminative features. Second, non-dermoscopic images (Figure 6 (b)), which differ significantly from the standard acquisition protocol and may interfere with the model's learning and predictive behavior. Although such cases represent a small fraction of the dataset, their presence may nonetheless affect both training stability and inference reliability.

5. Discussion, Limitations and Future Perspectives

The results obtained with MultiExCam show a weighted F1-score of 0.86, with precision and recall of 0.86 and 0.85, which are competitive compared to some related works, as [9]. Models such as SKINC-NET [5] or SkinDWNet+GB [8] achieve higher accuracies, above 97-98%, thanks to more complex architectures and advanced data augmentation strategies. However, these solutions require extensive preprocessing, a large number of parameters, and often lack explainability, limiting the transparency of their predictions. In various medical contexts, and beyond, we have shown how important it is not only to achieve good results but also to explain the reasoning behind them, in order to fully understand from a clinical perspective how reliable they are [14], as is already done in MultiExCam. However, this does not justify achieved performances and further investigation is needed. In fact, despite the encouraging results obtained, the current implementation of our framework presents some limitations that merit discussion. Firstly, the ensemble strategy followed the same protocol as the original MultiExCam implementation, operating on the discrete predictions of the individual models rather than on their

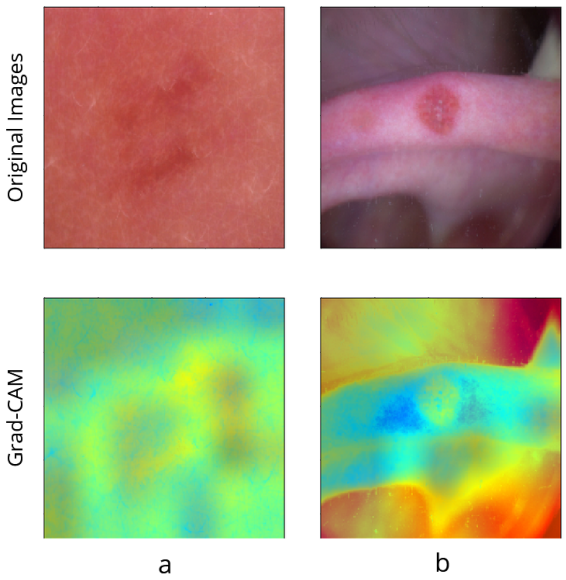

Figure 6: Examples of challenging images identified through Grad-CAM analysis: (a) non-dermoscopic image, (b) image with poorly visible lesion area.

predicted probabilities. Utilizing probabilistic outputs would likely provide a richer signal for the ensemble, potentially enhancing its discriminative capability, especially in multiclass scenarios [15]. Secondly, while both the machine learning models and the neural network were retrained, including hyperparameter tuning and architectural modifications to accommodate three classes, no task-specific strategies, such as a refined model selection or weighting scheme, were employed. This design choice was intentional, aiming to evaluate the robustness of the originally proposed solution in a new, more challenging multiclass context. Nevertheless, these decisions also expose clear areas of improvement; the reliance on discrete predictions limits the potential of the ensemble, and the lack of task-specific optimization suggests that performance could be further enhanced with a more targeted strategy. Given the new multi-class scenario, it is important to test different deep learning architectures, such as ConvNeXtV2 [7] or Vision Transformers [9], which have so far been evaluated only in binary classification tasks in our original framework. overall, while the current approach demonstrates the feasibility and generalization potential of the MultiExCam framework in an expanded multi-class setting, these limitations highlight several avenues for future development. These include probability-based ensemble methods, selective model inclusion, advanced data augmentation techniques, and the extension of the framework to cover additional classes of skin lesions beyond the three considered here. Addressing these aspects will not only improve classification performance but also strengthen the clinical applicability of MultiExCam, moving it closer to a comprehensive and interpretable tool for automated skin lesion diagnosis.

6. Conclusions

In this study, we have explored the application of MultiExCam framework in a multi-class skin lesion classification scenario, demonstrating that it can generalize beyond the originally tested setting while maintaining performance interpretability. Our results show that MultiExCam provides competitive performance on multiple lesion classes and offers explanations that support clinical understanding, addressing a key limitation of many high-performing yet opaque DL models.

At the same time, our findings highlight persistent challenges, particularly regarding the quality and diversity of available datasets. Skin lesion datasets remain imbalanced, with some classes under-represented, which can limit model performance and generalization. In addition, similar to what has been reported for the melanoma class, extending the same datasets to multi-class scenarios reveals the

same underlying bias, as most images still correspond to lighter skin phototypes, limiting the models' applicability across diverse populations [16, 17]. Addressing these limitations is crucial not only for improving accuracy across all lesion types but also for ensuring equitable clinical decision support.

Future work will focus on extending MultiExCam to additional lesion classes and testing advanced deep learning architectures, ensemble strategies, and sophisticated data augmentation techniques. Only by addressing both model and data limitations can we move toward robust, interpretable, and clinically relevant automated skin lesion diagnosis.

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat-GPT-4 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] Iarc. world health organization., 2024. https://gco.iarc.fr/.

[2] C. Garbe, T. Amaral, K. Peris, et al., European consensus-based interdisciplinary guideline for melanoma. part 1: Diagnostics: Update 2022, European Journal of Cancer 170 (2022) 236–255.

[3] M. Pippione, et al., Dermatologia e malattie sessualmente trasmissibili, Minerva medica, 2015.

[4] T. Ruga, L. Caroprese, E. Vocaturo, E. Zumpano, Multiexcam: A multi approach and explainable artificial intelligence architecture for skin lesion classification, Computer Methods and Programs in Biomedicine (2025) 109081.

[5] S. Asif, S. U. R. Khan, K. Amjad, M. Awais, Skinc-net: An efficient lightweight deep learning model for multiclass skin lesion classification in dermoscopic images, Multimedia Tools and Applications 84 (2025) 12531–12557.

[6] M. A. Khan, D. Rastogi, P. Johri, A. Al-Taani, V. S. Baghela, Kumud, Hybrid deep cnn model for multi-class classification of skin lesion, Neural Computing and Applications 37 (2025) 19479–19499.

[7] B. Ozdemir, I. Pacal, A robust deep learning framework for multiclass skin cancer classification, Scientific Reports 15 (2025) 4938.

[8] A. Naeem, H. Malik, M.-z. Din, A. Sadeghi-Niaraki, D. Jeong, R. A. Naqvi, Skindwnet: a novel deep learning model for multiclass classification of skin cancers using dermoscopic images, Multimedia Systems 31 (2025) 314.

[9] H. Erbay, Y. M. Abulgasim, D. E. Özer, F. Ertürk, Enhancing multi-class skin lesion diagnosis through ensemble learning of cnn and transformer architectures, Engineering Science and Technology, an International Journal 70 (2025) 102145.

[10] T. Ruga, G. Musacchio, D. Maurmo, An ensemble architecture for melanoma classification, in: pHealth 2024, IOS Press, 2024, pp. 183–184.

[11] P. Tschandl, C. Rosendahl, H. Kittler, The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Scientific data 5 (2018) 1–9.

[12] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al., Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), arXiv:1902.03368 (2019).

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE ICCV, 2017, pp. 618–626.

[14] T. Ruga, E. Vocaturo, E. Zumpano, Explainable deep learning for chest x-ray classification, in: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2024, pp. 6561–6566.

[15] N. Verma, Ranvijay, D. K. Yadav, A comprehensive review on step-based skin cancer detection

using machine learning and deep learning methods, Archives of Computational Methods in Engineering (2025) 1–54.

[16] T. Ruga, E. Zumpano, E. Vocaturo, L. Caroprese, C. Arlia, Bias in dermatological datasets: A critical analysis of the underrepresentation of dark skin tones in melanoma classification images, in: International Conference on Computational Science, Springer, 2025, pp. 434–448.

[17] T. Ruga, E. Zumpano, E. Vocaturo, Underrepresentation of dark skin tone in skin lesion datasets: The role of the explainable techniques in assessing the bias, in: New Trends in Database and Information Systems: ADBIS 2025 Short Papers, Workshops, Doctoral Consortium and Tutorials, Tampere, Finland, September 23–26, 2025, Proceedings, Springer Nature, 2025, p. 446.