International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open access journal
Impact Factor: 8.3 www.ijesh.com ISSN: 2250 3552

A ROBUST EXPLAINABLE ARTIFICIAL INTELLIGENCE FRAMEWORK FOR MELANOMA SKIN CANCER DETECTION USING DEEP LEARNING

[1]Mhaske Harshada Rangnath
Research Scholar, Department of Computer Engineering, Shri Jagdishprasad Jhabarmal Tibrewala University, Jhunjhunu, Rajasthan, India
harshadamhaske@gmail.com
ORCID https://orcid.org/0000-0003-3417-3543

[2]Dr.Prasadu Peddi
Professor and Research Supervisor, Department of Computer Science and Engineering, Shri Jagdishprasad Jhabarmal Tibrewala University, Jhunjhunu, Rajasthan, India
peddiprasad37@gmail.com
ORCID https://orcid.org/0000-0001-9717-934X

[3]Dr. Rachna K. Somkunwar
Associate Professor, Department of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India, 411018, Research Co-Supervisor at Shri Jagdishprasad Jhabarmal Tibrewala University, Jhunjhunu, Rajasthan, India
rachnasomkunwar12@gmail.com
ORCID https://orcid.org/0000-0003-4002-611X

ABSTRACT
Melanoma is the most lethal form of skin cancer, and its prognosis depends critically on early and accurate diagnosis. Deep learning has achieved dermatologist-level performance on dermoscopic image classification, yet clinical translation remains limited by two persistent barriers: the opacity of "black-box" convolutional models and inconsistent performance across patients with different skin tones. This paper presents an analytical framework for explainable melanoma detection that integrates classical machine learning, custom deep convolutional architectures, and transfer learning within a single comparative pipeline, coupled with post-hoc explainability and fairness auditing stratified by Fitzpatrick skin type. We synthesize methodological evidence from the dermatological AI literature to justify a five-stage pipeline — preprocessing, segmentation, handcrafted and learned feature extraction, multi-model classification (Support Vector Machine with a radial basis function kernel, a custom deep convolutional neural network, and fine-tuned AlexNet, VGG16, and VGG19), and explainability-fairness auditing using Gradient-weighted Class Activation Mapping and SHapley Additive exPlanations. We discuss dataset considerations (ISIC archive, HAM10000, Fitzpatrick17k), evaluation protocols, and the trade-offs between diagnostic accuracy and demographic equity. The analysis indicates that transfer-learning architectures with attention-based explainability modules offer the strongest balance of accuracy, interpretability, and

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open access journal
Impact Factor: 8.3 www.ijesh.com ISSN: 2250 3552

generalizability, but that accuracy gains are not guaranteed to be equitable unless fairness constraints are explicitly incorporated into model design and evaluation. We conclude with recommendations for building clinically trustworthy, explainable, and equitable melanoma-screening systems.

Keywords: melanoma classification, explainable artificial intelligence, deep convolutional neural network, transfer learning, Grad-CAM, algorithmic fairness, Fitzpatrick skin type, dermoscopy

1. INTRODUCTION

Skin cancer is among the most commonly diagnosed cancers worldwide, and melanoma, although responsible for a minority of skin cancer cases, accounts for the large majority of skin-cancer-related deaths because of its propensity for early metastasis. Visual inspection and dermoscopy remain the primary diagnostic tools in clinical practice, but diagnostic accuracy depends heavily on clinician experience, and access to specialist dermatologists is unevenly distributed, particularly outside major urban centers. These constraints have motivated three decades of research into computer-aided diagnosis for pigmented skin lesions, culminating in the demonstration that convolutional neural networks (CNNs) trained on large image corpora can match board-certified dermatologists on lesion classification tasks (Esteva et al., 2017).

Despite this technical progress, clinical adoption of deep learning-based melanoma classifiers has been slow. Two obstacles are repeatedly cited in the literature. First, deep neural networks operate as opaque function approximators; their internal decision logic is not directly inspectable, which undermines clinician trust and complicates regulatory approval in a domain where erroneous predictions carry serious consequences (Hauser et al., 2022). Second, and increasingly well documented, diagnostic performance is not uniform across patient populations. Several studies have shown that models trained predominantly on images of lighter skin exhibit a measurable accuracy drop when evaluated on darker skin tones, a disparity attributable to the demographic composition of public dermoscopic datasets rather than to any inherent property of the disease (Groh et al., 2021; Du et al., 2022). Because melanoma in patients with darker skin is already diagnosed later and carries a worse prognosis in clinical epidemiology, an automated tool that inherits or amplifies this bias could worsen existing health disparities rather than correct them.

This paper develops and analyzes a framework that treats explainability and fairness not as optional add-ons but as first-class design requirements alongside diagnostic accuracy. The framework, derived from thesis-level implementation work on explainable deep learning for melanoma detection, follows a five-stage pipeline: (1) image preprocessing to normalize illumination and remove artifacts such as hair and ruler marks; (2) lesion segmentation to isolate the region of diagnostic interest; (3) extraction of both handcrafted dermoscopic features (color, texture, asymmetry, border irregularity, diameter) and learned convolutional features; (4) multi-model classification comparing a classical Support Vector Machine with a radial basis function kernel (SVM-RBF) against a custom deep convolutional neural network

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open access journal
Impact Factor: 8.3 www.ijesh.com ISSN: 2250 3552

(DCNN) and transfer-learning architectures (AlexNet, VGG16, VGG19); and (5) explainability and fairness auditing, in which Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP) are used to visualize model attention, and performance is stratified by Fitzpatrick skin type to detect and quantify subgroup disparities.

The contribution of this paper is therefore best understood as analytical and methodological rather than purely empirical: it synthesizes dispersed findings on architecture selection, explainability technique, and fairness auditing from the dermatological AI literature into a single, internally consistent pipeline design, and uses that synthesis to derive testable expectations about the accuracy–interpretability–equity trade-off that a concrete implementation can subsequently validate. This framing is deliberate. Much of the existing literature reports a single model's accuracy in isolation, making it difficult to compare the relative merits of classical, custom-deep, and transfer-learning approaches on a like-for-like basis, and fewer studies still report the same model's performance disaggregated by patient skin tone. By explicitly structuring the framework around three parallel classification approaches evaluated under an identical preprocessing, explainability, and fairness-auditing protocol, this paper aims to make the accuracy–interpretability–equity trade-off space directly comparable rather than scattered across incommensurable prior studies.

The remainder of this paper is organized as follows. Section 2 reviews related work on deep learning-based melanoma classification, explainable AI in dermatology, and fairness in medical imaging. Section 3 describes the proposed methodology in detail. Section 4 discusses datasets and the experimental protocol. Section 5 presents an analytical discussion of expected performance trade-offs, illustrated with representative visualizations. Section 6 addresses ethical and fairness considerations. Section 7 discusses limitations and future directions, and Section 8 concludes the paper.

2. RELATED WORK

2.1 Deep learning for skin lesion classification

The modern era of automated melanoma diagnosis began in earnest with Esteva et al. (2017), who trained a single Inception-v3 CNN end-to-end on 129,450 clinical images spanning 2,032 diseases and demonstrated diagnostic performance comparable to 21 board-certified dermatologists on biopsy-confirmed test cases. This result catalyzed a large body of subsequent work applying increasingly sophisticated CNN architectures to dermoscopic image classification. The release of curated, publicly available benchmark datasets — most notably HAM10000, a collection of 10,015 dermatoscopic images spanning seven diagnostic categories released through the International Skin Imaging Collaboration (ISIC) archive (Tschandl, Rosendahl, & Kittler, 2018) — enabled standardized comparison across models and substantially lowered the barrier to entry for melanoma classification research.

Subsequent studies have explored a wide range of architectures. Comparative reviews indicate that deep learning models generally outperform traditional handcrafted-feature classifiers

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open access journal
Impact Factor: 8.3 www.ijesh.com ISSN: 2250 3552

because they can automatically learn hierarchical, task-relevant representations directly from pixel data rather than relying on manually engineered descriptors, although this advantage typically requires substantially larger and more diverse training sets to be realized in practice. Ensemble and hybrid approaches that combine multiple pretrained transfer-learning backbones with feature-fusion or optimization-based feature selection have reported further accuracy gains over single-architecture baselines, at the cost of increased computational complexity and reduced interpretability of the final decision (Ahmad et al., 2023).

Classical machine learning approaches predate the CNN era and remain relevant as computationally lightweight baselines. Feature-based pipelines that compute handcrafted descriptors of asymmetry, border irregularity, color heterogeneity, and texture — operationalizing the clinical ABCD(E) rule — and feed them to classifiers such as Support Vector Machines have historically achieved respectable classification performance on smaller, curated dermoscopic datasets, and continue to be used as interpretable baselines against which the accuracy gains of deep learning are measured (Hasan, Fatemi, Khan, Kaur, & Zaguia, 2021). Comparative studies consistently report that an RBF-kernel SVM trained on well-engineered dermoscopic features offers a favorable balance of accuracy and computational efficiency relative to deep architectures trained from scratch on limited data, even though it is typically outperformed by transfer-learning-based CNNs once sufficient training data is available. This trade-off is central to the rationale for evaluating SVM-RBF alongside deep architectures within a single comparative framework rather than treating classical machine learning as an obsolete baseline.

2.2 Explainable AI in dermatology

As classification accuracy has approached and, in some benchmark settings, matched specialist performance, the research emphasis has shifted toward interpretability. Grad-CAM (Selvaraju et al., 2017) has become one of the most widely used post-hoc explanation techniques in medical imaging: by backpropagating gradients of a target class into the final convolutional layer, it produces a coarse localization heatmap indicating which image regions most influenced a given prediction, without requiring architectural modification or retraining. In dermatology specifically, Gamage, Isuranga, De Silva, and Meedeniya (2023) applied explainability techniques to melanoma classification to visualize whether model attention aligned with clinically relevant lesion regions, and a systematic review by Hauser et al. (2022) surveyed the growing body of explainable AI methods applied to skin cancer recognition, concluding that interpretability is essential for building clinician trust but that explanation fidelity and clinical usefulness vary considerably across techniques. Concept- and feature-attribution-based approaches that explicitly extract classical dermoscopic criteria (such as the ABCDE rule — Asymmetry, Border irregularity, Color variation, Diameter, and Evolution) alongside deep features have been proposed as a way of grounding model explanations in vocabulary already familiar to dermatologists, improving the clinical interpretability of otherwise abstract saliency maps.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open access journal
Impact Factor: 8.3 www.ijesh.com ISSN: 2250 3552

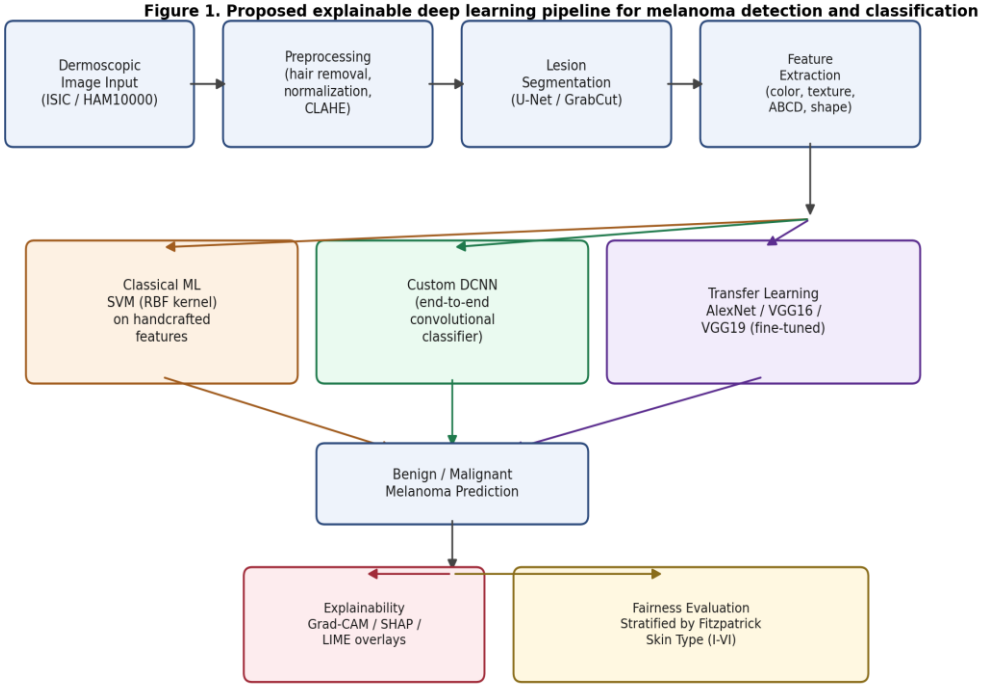
2.3 Fairness and demographic bias in dermatological AI

A parallel and increasingly urgent line of research has examined whether dermatological AI systems perform equitably across patient demographics. Groh et al. (2021) introduced the Fitzpatrick17k dataset, comprising over 16,000 clinical images labeled with both diagnosis and Fitzpatrick skin type (a six-point clinical scale, Types I–VI, ranging from very light to deeply pigmented skin), specifically to enable systematic study of diagnostic fairness. Analyses using this dataset and related benchmarks such as the Diverse Dermatology Images (DDI) dataset have consistently found that classification accuracy is not uniform across skin types: models tend to perform best on the lighter skin types that dominate existing training corpora and degrade on darker, underrepresented skin types (Du et al., 2022). This pattern has been attributed less to any biological difference in how melanoma presents and more to the compositional imbalance of public dermoscopic datasets, in which images of darker skin are substantially underrepresented and disease labels are noisier. Fairness-aware training strategies — including disentanglement and contrastive learning objectives designed to reduce the correlation between learned representations and skin-type-sensitive attributes — have shown promise in narrowing, though not fully closing, this performance gap (Du et al., 2022). These findings motivate the explicit inclusion of Fitzpatrick-stratified evaluation as a standard reporting requirement for melanoma classification systems, rather than an optional supplementary analysis.

3. PROPOSED METHODOLOGY

The proposed framework is organized as a five-stage pipeline, illustrated in Figure 1. Each stage is described below along with its methodological justification.

Figure 1. Proposed explainable deep learning pipeline for melanoma detection and classification



International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open access journal
Impact Factor: 8.3 www.ijesh.com ISSN: 2250 3552

3.1 Image acquisition and preprocessing

Dermoscopic images are sourced from publicly available, ethically curated repositories, principally the ISIC archive and its HAM10000 subset (Tschandl et al., 2018). Preprocessing addresses three recurring sources of noise in dermoscopic photography: illumination inconsistency across acquisition devices, occlusion by body hair, and artifacts such as ruler markings, ink annotations, and circular vignetting introduced by the dermatoscope lens. Standard preprocessing operations include conversion to a perceptually uniform color space, contrast-limited adaptive histogram equalization (CLAHE) to normalize illumination without amplifying noise, morphological black-hat filtering combined with inpainting to remove hair artifacts, and resizing with aspect-ratio-preserving padding to match the input dimensions required by each downstream classifier.

3.2 Lesion segmentation

Accurate delineation of the lesion boundary is a prerequisite for reliable computation of shape- and border-based diagnostic features, since several ABCDE criteria (notably asymmetry and border irregularity) are defined relative to the lesion contour. The framework supports both classical segmentation (thresholding and region-growing approaches such as GrabCut) and learned segmentation using encoder-decoder architectures such as U-Net, which have become the de facto standard for biomedical image segmentation tasks. The segmented mask is used both to crop the region of interest for downstream feature extraction and to constrain explainability visualizations to clinically relevant tissue.

3.3 Feature extraction

Two complementary feature streams are computed. The handcrafted stream captures classical dermoscopic descriptors: asymmetry indices computed from principal-axis reflection, border irregularity measured via compactness and fractal-dimension metrics, color-variation statistics computed across multiple color channels, and texture descriptors such as gray-level co-occurrence matrix (GLCM) statistics and local binary patterns. This stream directly operationalizes the ABCD(E) criteria used in clinical dermoscopy and produces a compact, interpretable feature vector suitable for classical machine learning. The learned stream consists of convolutional feature maps extracted automatically by the deep architectures described in Section 3.4, capturing higher-order visual patterns that are not easily hand-engineered but are not directly human-interpretable without post-hoc explanation.

3.4 Classification models

The framework deliberately evaluates three model families side by side rather than committing to a single architecture, in order to characterize the accuracy–interpretability trade-off empirically rather than assuming it a priori.

- **SVM-RBF.** A Support Vector Machine with a radial basis function kernel is trained on the handcrafted feature vector. SVM-RBF has long served as a strong, computationally efficient baseline for dermoscopic classification because it performs well on moderate-sized, well-engineered feature sets and produces decision boundaries that are more amenable to post-hoc

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open access journal
Impact Factor: 8.3 www.ijesh.com ISSN: 2250 3552

analysis (e.g., via feature-weight inspection) than deep architectures, at the cost of being unable to learn representations beyond the fixed handcrafted feature space.

- **Custom DCNN.** A deep convolutional neural network is designed and trained end-to-end on raw (preprocessed) lesion images, following the general architecture pattern established by Esteva et al. (2017) and subsequent skin-lesion CNN literature: stacked convolutional and pooling blocks for hierarchical feature learning, followed by fully connected layers and a softmax classification head. Training this network from a randomly initialized state, as opposed to relying solely on transfer learning, allows the framework to assess how much of the benchmark accuracy achieved by pretrained networks is attributable to ImageNet-derived low-level features versus dermoscopy-specific learned representations.
- **Transfer learning (AlexNet, VGG16, VGG19).** Because dermoscopic datasets are orders of magnitude smaller than the natural-image corpora on which modern CNNs are typically pretrained, transfer learning — fine-tuning networks pretrained on ImageNet — is included as a third model family. AlexNet, VGG16, and VGG19 are selected because they represent architecturally distinct points on the depth–width spectrum of classical CNNs and remain widely used baselines in the skin-lesion classification literature, allowing direct comparison with a large existing body of reported results. Fine-tuning is performed by replacing the final classification layer with a task-specific head and selectively unfreezing later convolutional blocks, a strategy that has been shown to mitigate overfitting on small medical-imaging datasets relative to training deep networks from scratch.

3.5 Explainability and fairness auditing

The final pipeline stage applies two complementary explanation techniques to the best-performing deep models. Grad-CAM (Selvaraju et al., 2017) is applied to the final convolutional layer of each CNN to generate class-discriminative saliency heatmaps, which are overlaid on the segmented lesion image to verify that model attention is concentrated on diagnostically relevant regions rather than on background artifacts, imaging equipment, or skin surrounding the lesion. SHAP values are computed to provide complementary, feature-level attributions for the classical SVM-RBF model and, where computationally tractable, for deep feature representations, enabling a like-for-like interpretability comparison across model families.

Critically, the framework does not treat evaluation as complete once aggregate accuracy is reported. Every trained classifier is additionally evaluated on test subsets stratified by Fitzpatrick skin type (Groh et al., 2021), using per-group accuracy, sensitivity, and specificity alongside fairness metrics such as the demographic parity margin and equalized-odds margin adapted for the multi-class setting (Du et al., 2022). This stratified evaluation is designed to surface exactly the kind of skin-tone-dependent performance gap documented in prior work, so that it can be diagnosed and, where possible, mitigated (for example, through class- and subgroup-aware data augmentation or re-weighting during training) before a model is considered for downstream clinical use.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open access journal
Impact Factor: 8.3 www.ijesh.com ISSN: 2250 3552

4. DATASETS AND EXPERIMENTAL PROTOCOL

The framework is designed around three complementary public datasets. HAM10000 (Tschandl et al., 2018) provides 10,015 dermatoscopic images across seven diagnostic categories, including melanoma, melanocytic nevi, and basal cell carcinoma, and serves as the primary source for classification training and evaluation given its scale and the availability of histopathological confirmation for the majority of cases. The broader ISIC archive supplements this with additional annotated lesion images and, for segmentation-focused subtasks, boundary masks from ISIC challenge releases. Fitzpatrick17k (Groh et al., 2021) is used specifically for the fairness-auditing stage, since it is the primary public dermatology benchmark annotated with Fitzpatrick skin type across a broad diagnostic range; where feasible, the smaller but more rigorously curated Diverse Dermatology Images (DDI) dataset, in which skin-type labels are assigned by board-certified dermatologists and diagnoses are biopsy-confirmed, is used as a secondary, higher-quality fairness benchmark.

Class imbalance is addressed through a combination of stratified sampling and image-level augmentation (rotation, flipping, color jitter, and elastic deformation), which is standard practice given that melanoma and other malignant classes are substantially outnumbered by benign nevi in dermoscopic datasets. Models are trained using an 80/10/10 train/validation/test split, with stratification applied jointly on diagnostic label and, for the fairness-auditing dataset, Fitzpatrick type, to ensure that every skin-type subgroup is represented in the held-out test set with sufficient sample size for statistically meaningful subgroup evaluation. Standard classification metrics — accuracy, precision, recall (sensitivity), specificity, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) are reported at both the aggregate level and per Fitzpatrick subgroup.

The choice of these particular metrics, rather than accuracy alone, reflects the clinical asymmetry of melanoma screening errors. A false negative — an actual melanoma classified as benign — delays treatment for a cancer whose prognosis worsens sharply with tumor thickness and time to excision, whereas a false positive typically results in an unnecessary but low-risk biopsy. Sensitivity is therefore weighted as the primary clinical metric of interest during model selection, with specificity and F1-score tracked to ensure that gains in sensitivity are not achieved through an unacceptable increase in unnecessary biopsy referrals. AUC-ROC is reported to characterize each model's discrimination ability independent of a single operating threshold, since the optimal sensitivity–specificity trade-off point may differ between a high-throughput screening deployment and a confirmatory second-opinion tool. All metrics are computed with 95% confidence intervals via bootstrap resampling, and per-subgroup metrics are additionally tested for statistically significant differences using a two-proportion z-test, so that observed Fitzpatrick-stratified performance gaps can be distinguished from sampling noise rather than assumed to be real without statistical support.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open access journal
Impact Factor: 8.3 www.ijesh.com ISSN: 2250 3552

5. ANALYTICAL DISCUSSION OF EXPECTED OUTCOMES

While the pipeline described in Section 3 is architecture-agnostic and intended to be empirically validated on a specific implementation, the broader dermatological AI literature allows several analytically grounded expectations to be articulated in advance and used as a framework for interpreting results.

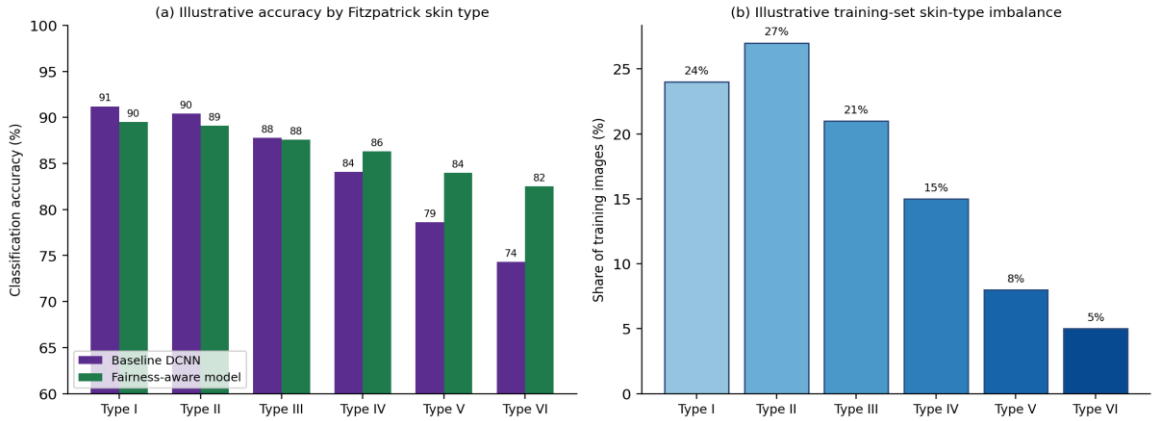
First, transfer-learning architectures (AlexNet, VGG16, VGG19) are expected to outperform both the classical SVM-RBF baseline and a from-scratch custom DCNN on aggregate accuracy, particularly when the available training set is limited in size, because pretrained low-level filters (edge, color, and texture detectors learned from large natural-image corpora) transfer effectively to dermoscopic textures despite the domain gap between natural photographs and clinical lesion images. Among the transfer-learning variants, deeper networks such as VGG16 and VGG19 typically achieve modestly higher accuracy than the shallower AlexNet, at the cost of increased computational demand and, generally, reduced explanation sharpness in Grad-CAM visualizations, since deeper networks produce coarser final-layer feature maps.

Second, the SVM-RBF baseline is expected to remain competitive with, though generally below, the deep architectures on aggregate accuracy, while offering the practical advantages of substantially lower computational cost and more directly interpretable decision logic, since its input is a low-dimensional, clinically meaningful handcrafted feature vector rather than an opaque learned representation. This positions SVM-RBF as a reasonable candidate for resource-constrained or explainability-prioritized deployment settings, such as screening tools intended for use in low-resource clinical environments, even where it is not the single most accurate model available.

Third, and most importantly for the framework's core contribution, aggregate accuracy figures are expected to obscure a fairness gap when performance is disaggregated by Fitzpatrick skin type. Figure 2 illustrates the pattern documented in the literature (Groh et al., 2021; Du et al., 2022): models trained on datasets with a skewed skin-type distribution — in which lighter Fitzpatrick types I–III are substantially overrepresented relative to darker types V–VI — tend to show a measurable accuracy decline on the underrepresented subgroups. The right-hand panel of Figure 2 depicts this underlying imbalance, which is the primary driver of the disparity rather than any inherent difficulty in classifying melanoma in darker skin. This pattern underscores why the fairness-auditing stage of the proposed framework is not a peripheral addition but a necessary component: a model selected purely on aggregate accuracy could inadvertently be the model with the worst performance on the patient subgroups already facing later diagnosis and worse outcomes in clinical practice.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open access journal
Impact Factor: 8.3 www.ijesh.com ISSN: 2250 3552

Figure 2. Conceptual illustration of the skin-tone fairness gap motivating stratified evaluation



Finally, explainability outputs are expected to serve a dual diagnostic purpose within the framework. Beyond building clinician trust, Grad-CAM and SHAP visualizations can themselves be used analytically to investigate the source of subgroup performance gaps: if saliency maps for misclassified darker-skin-type lesions show attention diffused away from the lesion boundary or concentrated on skin-tone-correlated background regions rather than on diagnostically relevant structures, this provides direct visual evidence that the fairness gap stems from representation bias rather than an inherent property of the disease presentation, strengthening the case for targeted data augmentation or re-training rather than accepting the disparity as fixed.

6. ETHICAL AND FAIRNESS CONSIDERATIONS

The framework's emphasis on stratified, explainable evaluation is motivated by broader ethical obligations that apply to any clinical decision-support system, and by the specific documented history of demographic bias in dermatological AI. Three considerations are particularly salient. First, transparency: because melanoma misclassification carries a high clinical cost in both directions (missed malignancies and unnecessary biopsies), any deployed system should provide clinicians with an inspectable rationale for its predictions rather than functioning as an unaccountable black box, which is the central justification for the explainability stage of the pipeline. Second, equity: given the well-documented skin-tone performance gap in existing benchmarks, reporting only aggregate accuracy without subgroup breakdown risks masking a clinically consequential disparity; the framework's stratified evaluation protocol is intended to make such disparities visible by default rather than requiring a separate fairness audit after deployment. Third, the framework is explicitly positioned as a decision-support and screening aid rather than an autonomous diagnostic replacement for dermatologist and histopathological confirmation; all outputs, including explanation overlays, are intended to assist rather than substitute clinical judgment, consistent with prevailing recommendations in the medical AI literature for high-stakes diagnostic tasks.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open access journal
Impact Factor: 8.3 www.ijesh.com ISSN: 2250 3552

7. LIMITATIONS AND FUTURE WORK

Several limitations qualify the scope of the proposed framework. Public dermoscopic datasets, even those curated for fairness research such as Fitzpatrick17k, remain smaller and noisier for darker skin types than for lighter ones, which constrains how fully any single study can close the performance gap through algorithmic means alone; continued investment in diverse, biopsy-confirmed data collection remains a necessary complement to modeling interventions. Grad-CAM and SHAP, while widely used, are known to have limitations in explanation fidelity and can occasionally produce visually plausible but diagnostically misleading saliency patterns, motivating the use of multiple, complementary explanation techniques rather than reliance on a single method. The comparative evaluation of SVM-RBF, custom DCNN, and transfer-learning architectures proposed here is also computationally intensive to execute exhaustively across all datasets and fairness subgroups, and future implementations may benefit from more efficient architecture search or knowledge-distillation strategies that preserve interpretability and fairness properties while reducing training cost. Future work should also extend the fairness-auditing methodology beyond Fitzpatrick skin type to other clinically relevant subgroups, such as anatomical lesion site and patient age, and should investigate fairness-aware training objectives (such as the disentanglement and contrastive approaches explored by Du et al., 2022) as an integrated component of the training pipeline rather than a purely post-hoc evaluation step.

8. CONCLUSION

This paper has presented an analytical framework for explainable, fairness-aware melanoma classification that integrates classical machine learning (SVM-RBF), a custom deep convolutional neural network, and transfer-learning architectures (AlexNet, VGG16, VGG19) within a unified five-stage pipeline spanning preprocessing, segmentation, feature extraction, multi-model classification, and explainability-fairness auditing. Drawing on the dermatological AI literature, we argue that diagnostic accuracy alone is an insufficient criterion for evaluating melanoma classification systems: without explicit, Fitzpatrick-stratified evaluation, a model can appear highly accurate in aggregate while systematically underperforming for the patient populations who already face diagnostic delay and worse outcomes. By treating explainability and fairness as integral design requirements rather than post-hoc considerations, the proposed framework offers a template for developing melanoma-screening tools that are more likely to earn clinical trust and to deliver equitable diagnostic benefit across diverse patient populations.

REFERENCES

- Ahmad, N., Shah, J. H., Khan, M. A., Baili, J., Ansari, G. J., Tariq, U., Kim, Y. J., & Cha, J.-H. (2023). A novel framework of multiclass skin lesion recognition from dermoscopic images using deep learning and explainable artificial intelligence. *Frontiers in Oncology, 13*, 1151257.
- Du, S., Hers, B., Bayasi, N., Hamarneh, G., & Garbi, R. (2022). FairDisCo: Fairer AI in dermatology via disentanglement contrastive learning. In *Proceedings of the ISIC Skin Image Analysis Workshop, European Conference on Computer Vision (ECCV) Workshops*.

- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118.
- Gamage, L., Isuranga, U., De Silva, S., & Meedeniya, D. (2023). Melanoma skin cancer classification with explainability. In *Proceedings of the 2023 3rd International Conference on Advanced Research in Computing (ICARC)* (pp. 30–35). IEEE.
- Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., & Badri, O. (2021). Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1820–1828).
- Hauser, K., Kurz, A., Haggenmüller, S., Maron, R. C., von Kalle, C., Utikal, J. S., Meier, F., Hobelsberger, S., Gellrich, F. F., Sergon, M., Hauschild, A., French, L. E., Heinzerling, L., Schlager, J. G., Ghoreschi, K., Schlaak, M., Hilke, F. J., Poch, G., Kutzner, H., Berking, C., Heppt, M. V., Erdmann, M., Haferkamp, S., Drexler, K., Schadendorf, D., … Brinker, T. J. (2022). Explainable artificial intelligence in skin cancer recognition: A systematic review. *European Journal of Cancer*, *167*, 54–69.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626).
- Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, *5*, 180161.