

A Hybrid Approach for Skin Lesion Classification Combining Handcrafted and Deep Learning Features

Soud Asaad

Artificial Intelligence Research Center (AIRC)
College of Engineering and IT, Ajman University
Ajman, UAE
Soudkhaled15@outlook.com

Mohamed Deriche

Artificial Intelligence Research Center (AIRC)
College of Engineering and IT, Ajman University
Ajman, UAE
m.deriche@ajman.ac.ae

Abstract—Accurate melanoma detection continues to be a major challenge in dermoscopy-based diagnosis. This paper proposes a hybrid framework that integrates handcrafted dermatological descriptors with transformation-based supervised deep features derived from a lightly fine-tuned CNN. The deep embeddings are transformed and dimension-reduced via Linear Discriminant Analysis (LDA) to produce a small set of highly discriminative components, which are then fused with the handcrafted shape, color, and texture descriptors to form a 39-dimensional feature representation. Classical machine learning classifiers, particularly LightGBM, were then applied for melanoma versus non-melanoma classification on the HAM10000 dataset. The proposed framework is shown to achieve a ROC-AUC of 0.958 and accuracy of 0.944, outperforming individual feature sets as well as several recently reported deep learning baselines. Our results highlight that careful feature engineering and feature fusion, combined with even basic machine learning methods, can achieve state-of-the-art diagnostic performance while remaining lightweight and computationally practical for clinical deployment.

Index Terms—Melanoma detection, dermoscopy, feature fusion, ResNet-50, LDA, ROC-AUC.

I. INTRODUCTION

Early detection of melanoma is critical given its high mortality rate relative to other cancers. Hence, timely diagnosis is crucial and improves survival chances, yet screening still remains challenging at scale [1]–[3]. Dermoscopy improves diagnostic accuracy over naked-eye inspection but demands expertise and remains variable across clinical settings [4], [5]. When using algorithmic approaches on skin images to detect cancer, we are faced with a number of challenges as well: what type of preprocessing is needed? what are the best representative features that could be extracted from such images, and finally what are the most appropriate classifiers for this problem. Feature design in computer-aided diagnosis (CAD) has traditionally followed two complementary tracks: (i) *domain-informed* descriptors inspired by dermatological rules (e.g., asymmetry, border irregularity, color variegation, texture) and (ii) *data-driven* representations learned from convolutional neural networks (CNNs). Classic handcrafted feature families (asymmetry, border, color, texture; ABCD)

encode clinically interpretable cues and are well established in the literature [6], [7], while CNN features capture high-level latent patterns that often surpass manual descriptors on large datasets.

However, real-world dermoscopy poses two persistent constraints: class imbalance (melanoma is rare) and limited labeled data, which can hinder end-to-end training and generalization. In this context, compact, discriminative representations, especially when fusing complementary sources, are attractive. Prior multimodal works on non-dermoscopic *clinical* images suggest that combining feature types (CNN plus handcrafted, optionally with metadata) can outperform single-source baselines [8], [9]; related conclusions highlight that clinical images generally carry less visual detail than dermoscopic images, further motivating careful representation design [10].

In this paper, We show that carefully engineered fusion of handcrafted dermatology features (33 shape–color–texture descriptors) with a set of LDA-based transformed supervised deep features can lead to excellent melanoma vs. non-melanoma classification results over the HAM10000 dataset without large end-to-end networks. In particular, we evaluate five classical classifiers (SVM, KNN, RF, XGBoost, LightGBM) across: handcrafted-only, 33+PCA (10/15/20), LDA(6)-only, and fusion (33+6=39). Our best model attains an excellent performance in AUC of = 0.958 with 39 features, outperforming recently reported results (typically AUC ~0.85–0.92). This supports our claim that smart feature engineering plus traditional ML algorithms can rival heavy CNNs for dermoscopy, offering efficiency and interpretability.

II. RELATED WORK

The literature on automated skin cancer detection spans handcrafted features with classical machine learning, deep learning with CNN-based extractors, and hybrid pipelines. While CNNs have achieved dermatologist-level accuracy, they require large datasets and heavy computation (i.e., high parameter counts and training requirements). Recent studies explore attention mechanisms, feature aggregation, and hybrid designs

that fuse handcrafted and deep features, aiming to balance accuracy, efficiency, and interpretability.

Hameed *et al.* [14] introduced a multi-level classification scheme using color and texture descriptors with SVMs, then deeper models for ambiguous instances. On ISIC datasets, this hierarchy achieved nearly 96% multiclass accuracy. The design manages lesions of differing complexity but relies on hand-crafted thresholds and does not report a binary melanoma benchmark, limiting direct comparison.

Tan *et al.* [15] paired improved particle swarm optimization (PSO) with ensemble classifiers over GLRLM, LBP, HOG, and ABCD-inspired features. On PH2 and ISIC 2016, accuracies reached 93.4% and 90.5%, with AUCs near 0.94. Alfed and Khelifi [16] combined textural and colour descriptors in a bag-of-features framework, reporting up to 98.4% accuracy and AUC of 0.98. Both approaches highlight discriminative subsets but depend on engineered features and raise computational cost.

In [11], the authors used an ImageNet-pretrained AlexNet as feature extractor on Dermofit, obtaining 81.8% accuracy for ten-class classification and 85.8% for five classes. Optimal feature extraction techniques have also been proposed [12]. Esteva *et al.* [24] trained an Inception v3 network, a deep model with millions of parameters, on 129k mixed images, reporting AUCs of 0.91–0.96. These results show CNNs deliver dermatologist-level accuracy but require large-scale data and heavy computation.

In [25] and [18], authors encoded CNN descriptors with Fisher vectors, improving ISIC 2017 accuracy to 85.5%. Zhang *et al.* [25] proposed attention residual learning, achieving 87.2% accuracy and 0.912 AUC. Attention improves focus on salient regions but increases training cost. codebook-based learning models have also been used.

Recent approaches span from handcrafted pipelines to hybrid CNN–feature fusion and attention-based networks. Table I summarizes representative methods and results.

TABLE I: Summary of Related Work

Study	Method	Dataset	Performance
Hameed [14]	SVM hierarchy	ISIC	Acc: 96%
Tan [15]	PSO + ensembles	PH2, ISIC	Acc: 90–93%, AUC: 0.94
Alfed [16]	Bag-of-features	PH2, Edinburgh	Acc: 93–98%, AUC: 0.98
Kawahara [11]	AlexNet + log. reg.	Dermofit	Acc: 81.8%
Esteva [24]	Inception v3 CNN	Clinical + dermoscopy	AUC: 0.91–0.96
Yu [17]	CNN + Fisher enc.	ISIC 2017	Acc: 85.5%
Zhang [25]	Attention ResNet	ISIC 2017	Acc: 87.2%, AUC: 0.912

III. THE PROPOSED HYBRID APPROACH

Based on our analysis of the literature and given the strengths of hybrid approaches, we propose our hybrid *feature-fusion* design combining dermatology-informed handcrafted

descriptors with supervised LDA-based features extracted from fine-tuned ResNet-50 embeddings. Unlike heavy end-to-end CNNs, our pipeline is small, interpretable, and attains top-tier ROC-AUC on the HAM10000 (melanoma vs. non-melanoma), while maintaining classical ML classifiers and transparent features. Figure 1 illustrates the proposed pipeline.

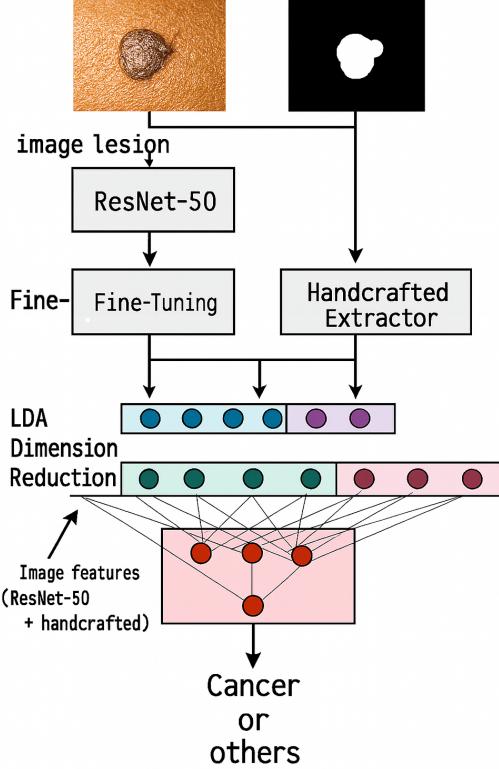


Fig. 1: Proposed Work Flow

A. Dataset Lesion Cropping and Preprocessing

We use the HAM10000 (10,015 dermoscopic images) and isolate the lesion region using binary masks to restrict measurements [19]. The train/test split preserves the test distribution; training images are augmented (random rotations, flips, shifts). Tabular features are standardized with StandardScaler; where supported we set `class_weight=balanced`.

To ensure features describe the lesion rather than background skin, we apply the binary masks to crop/weight the lesion area prior to measurement. Figure 2 shows an example of original dermoscopic images and the masks.

Handcrafted descriptors encode dermatology priors (shape, color, texture), while deep embeddings capture higher-level patterns. Fusing them yields complementary evidence with far fewer parameters than end-to-end CNNs; the supervised LDA stage further compresses deep features into a compact, discriminative subspace that works well with classical ML.

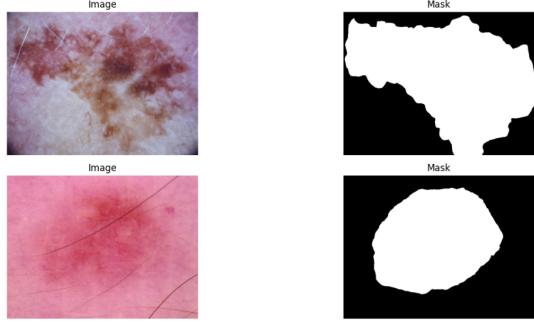


Fig. 2: Examples of original dermoscopic images on left side from HAM10000 dataset and their corresponding binary segmentation masks on right side.

B. Handcrafted Dermatology Features (33-D)

From the masked lesion regions, we extracted 33 descriptors grouped into three categories: shape (geometric irregularities such as area, perimeter, circularity, and asymmetry), texture (GLCM, LBP, entropy, and related statistics capturing pigmentation patterns), and color/intensity (RGB channel means/variances and grayscale statistics). These interpretable features reflect established dermatology heuristics and complement deep representations by providing transparent diagnostic cues.

C. Fine-Tuned Deep Features

We used ResNet-50 pretrained on ImageNet, with inputs resized to 224×224 , normalized, and augmented (rotations, flips, shifts). A lightweight head (*GlobalAveragePooling2D* \rightarrow *Dense(1, sigmoid)*) was added, and the network fine-tuned for 5 epochs with Adam ($\eta = 10^{-5}$) and binary cross-entropy (batch size 32). After training, we discarded the sigmoid and extracted the penultimate pooled activations as the deep descriptor.

To reduce dimensionality, we applied Linear Discriminant Analysis (LDA) trained on the original diagnostic labels. Unlike PCA, LDA leverages class information to maximize inter-class separation while minimizing intra-class variability, yielding a compact and interpretable space. Since our dataset has seven classes, LDA reduces the feature vector to six components. These LDA-based deep features integrate with handcrafted descriptors, creating a 39-D representation that is both discriminative and efficient.

D. Feature Fusion and Classification

We concatenate the handcrafted vector (33-D) with the LDA vector (6-D) into a 39-D feature vector. We evaluate SVM, KNN, Random Forest, XGBoost, and LightGBM with grid-search under stratified CV and report Accuracy, Precision, Recall, F1, and ROC-AUC. This fused representation delivers our best performance (ROC-AUC up to 0.958), outperforming either feature sets when used alone.

IV. EXPERIMENTS AND RESULTS

A. Feature Sets

For a comprehensive evaluation of the proposed framework, We considered several scenarios: handcrafted features only; Handcrafted with 10-D, 15-D, and 20-D PCA features from the non-fine-tuned ResNet-50 embeddings; FT-ResNet-50 (5 epochs, low LR, augmentation) \rightarrow deep embeddings \rightarrow LDA(6) for multi-class; LDA(6) only; fusion (33+6=9).

a) 33 handcrafted: Tree ensembles lead; LGBM gives best Accuracy (0.9066) and XGB best AUC (0.8826). This is a strong tabular baseline.

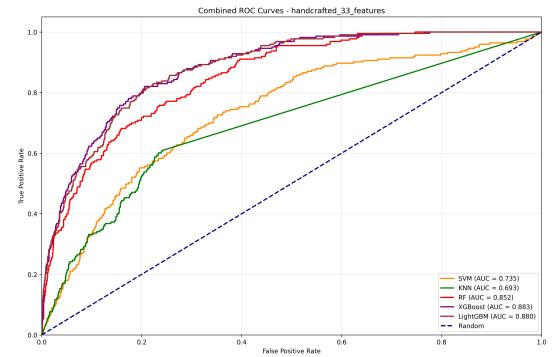


Fig. 3: E1: Combined ROC curves for the 33 handcrafted features (all classifiers).

b) 33 + PCA(10/15/20): Fusing PCA-reduced deep features consistently improves AUC; best at 20 comps (XGB 0.9174 AUC; LGBM 0.9146 Acc).

c) LDA(6) only: Despite being just 6-D, LGBM attains AUC 0.9020 and RF reaches 0.9071 Accuracy—showing LDA captures highly discriminative deep axes.

d) Fusion (33 + LDA(6) = 39-D): Feature fusion provides the largest gains: **LGBM achieves AUC 0.9580 and Acc 0.9436**, surpassing E1, E2, and E3.

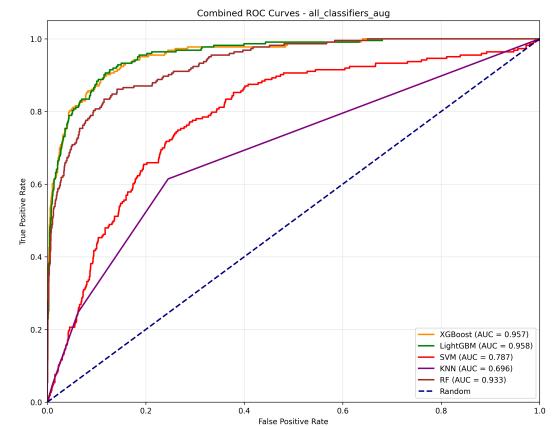


Fig. 4: E4 (33+LDA(6), 39-D fusion): Combined ROC curves (all classifiers).

B. Classifiers and Metrics

For each experiment, we train SVM, KNN, Random Forest, XGBoost, LightGBM with grid-search CV to fit best hyperparameters for each model. We report Accuracy, Precision, Recall, F1, ROC–AUC. Combined ROC curves are shown below in Table II.

TABLE II: Summary of Best Test-Set Performance per Experiment

Experiment (E1–E4)	Best Test-Set Performance		
	Best Classifier	Accuracy	ROC-AUC
E1: 33 handcrafted	LGBM / XGB	0.9066	0.8826
E2: 33 + PCA(10/15/20)	LGBM / XGB	0.9146	0.9174
E3: LDA(6) only	RF / LGBM	0.9071	0.9020
E4: 33 + LDA(6) (39-D)	LGBM	0.9436	0.9580

^a“Best Classifier” lists the top Accuracy / top ROC–AUC when they differ.

The boost comes from how these two sets of features work together: the handcrafted ones pick up on texture and color quirks that doctors can actually make sense of, while the tuned-up deep learning grab those subtler, big-picture patterns. Mixing them gives you a fuller picture of the lesion, striking a nice balance between being easy to understand and really effective at spotting differences.

V. COMPARISON WITH PREVIOUS RESEARCH

Recent melanoma detection studies on HAM10000 and related datasets typically report ROC–AUC values between 0.85 and 0.92. Esteva *et al.* achieved dermatologist-level performance with Inception v3 (AUC ~0.91–0.96), while Daneshjou *et al.* noted that baseline CNNs on HAM10000 reach 0.86–0.90. Larger architectures such as RegNetY-320 achieve ~0.95 but with tens of millions of parameters, and multimodal transformers exceed 0.92 at high computational cost. It is important to note that while some studies from our literature review (e.g., Alfed *et al.* [16]) reported very high AUCs (up to 0.98), those results were achieved on different datasets (e.g., PH2, Edinburgh) and are therefore not directly comparable to our performance on the HAM10000 benchmark.

Our fused 39-dimensional representation (33 handcrafted + 6 LDA features) attains **ROC–AUC = 0.958** with LightGBM, surpassing most CNN baselines while remaining lightweight and interpretable. Unlike deep networks requiring GPUs, our pipeline trains in minutes on a CPU, showing that feature fusion plus classical ML can rival or outperform heavy models.

TABLE III: Comparison of Binary Melanoma Classification Methods

Study	Method	AUC
Esteva et al. [20]	Inception v3	0.91–0.96
Daneshjou et al. [21]	CNN baselines	0.86–0.90
Ahsan et al. [22]	Optimized CNN	0.861
RegNetY-320 [22]	Deep CNN	0.95
Zhao et al. [23]	Multimodal transf.	>0.92
This work	Fusion (39-D)	0.958

VI. CONCLUSION

This fusion allows the model to ground the abstract patterns from the CNN with the clinically relevant, interpretable hand-crafted features, preventing overfitting to spurious correlations and improving generalization. We introduced, in this paper, a hybrid feature-fusion framework for early melanoma detection that combines handcrafted dermatology-inspired descriptors with LDA-compressed deep features extracted from a fine-tuned ResNet-50. The resulting 39-dimensional representation demonstrated superior performance on the HAM10000 dataset, achieving a ROC–AUC of 0.958 and accuracy of 0.944 using LightGBM. Compared to recent CNN-based and multimodal approaches, our method delivers competitive or better performance with far fewer parameters and reduced computational load. Importantly, the inclusion of handcrafted descriptors preserves interpretability, providing diagnostic interpretation aligned with established clinical heuristics. These results show that combining domain knowledge with supervised deep embeddings can even outperform heavy end-to-end neural networks. Future work will explore extending this framework to multiclass lesion classification, incorporating clinical metadata, and utilizing explainable AI (XAI). Future developments may integrate transformer-based or attention-guided fusion modules to facilitate more complex interactions between handmade and deep representations, potentially enhancing generalization across various lesion types.

VII. ACKNOWLEDGMENTS

This work has been partially funded by the Deanship of Research and Graduate Studies at Ajman University under Project 2025-IDG-CEIT-4.

REFERENCES

- [1] World Health Organization, “Radiation: Ultraviolet (UV) radiation and skin cancer,” [Online]. Available: [https://www.who.int/news-room/questions-and-answers/item/radiation-ultraviolet-\(uv\)-radiation-and-skin-cancer](https://www.who.int/news-room/questions-and-answers/item/radiation-ultraviolet-(uv)-radiation-and-skin-cancer). [Accessed: 20-Sep-2025].
- [2] R. S. Stern, “Prevalence of a history of skin cancer in 2007: results of an incidence-based model,” Arch Dermatol, vol. 146, no. 3, pp. 279–282, 2010.
- [3] World Cancer Research Fund, “Skin cancer statistics,” [Online]. Available: <https://www.wcrf.org/preventing-cancer/cancer-statistics/skin-cancer-statistics/>. [Accessed: 20-Sep-2025].
- [4] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, “Diagnostic accuracy of dermoscopy,” Lancet Oncol, vol. 3, no. 3, pp. 159–165, 2002.
- [5] M. Binder, M. Schwarz, A. Winkler, A. Steiner, A. Kaider, K. Wolff, and H. Pehamberger, “Epiluminescence microscopy: a useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists,” Arch Dermatol, vol. 131, no. 3, pp. 286–291, 1995.
- [6] F. Nachbar et al., “The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions,” J Am Acad Dermatol, vol. 30, no. 4, pp. 551–559, 1994.
- [7] G. Argenziano et al., “Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis,” Arch Dermatol, vol. 134, no. 12, pp. 1563–1570, 1998.
- [8] A. Pacheco and R. A. Krohling, “The impact of patient clinical information on automated skin cancer detection,” Comput Biol Med, vol. 116, 103545, 2020.
- [9] W. Li, J. Zhuang, R. Wang, J. Zhang, and W. Zheng, “Fusing metadata and dermoscopy images for skin disease diagnosis,” in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 1996–2000.

- [10] A. G. Pacheco et al., "PAD-UFES-20: a skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data in Brief*, vol. 32, 106221, 2020.
- [11] J. Kawahara, A. BenTaieb, and G. Hamarneh, "Deep features to classify skin lesions," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 2016, pp. 1397–1400.
- [12] A. Al-ani, M. Deriche, "Feature selection using a mutual information based measure," in *Proceedings of the International Conference on Pattern Recognition*, vol. 16, no. 4, pp. 82-85, 2002.
- [13] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.
- [14] N. Hameed, A. M. Shabut, M. K. Ghosh, and M. Hossain, "Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques," *Expert Syst Appl*, vol. 141, 112961, 2020.
- [15] T. Y. Tan, L. Zhang, and C. P. Lim, "Intelligent skin cancer diagnosis using improved particle swarm optimization and deep learning models," *Appl Soft Comput*, vol. 84, 105725, 2019.
- [16] N. Alfed and F. Khelifi, "Bagged textural and color features for melanoma skin cancer detection in dermoscopic and standard images," *Expert Syst Appl*, vol. 90, pp. 101–110, 2017.
- [17] Z. Yu et al., "Convolutional descriptors aggregation via cross-net for skin lesion recognition," *Appl Soft Comput*, vol. 92, 106281, 2020.
- [18] A. Amin, M. Deriche, "Salt-dome detection using a codebook-based learning model," *IEEE Geoscience and Remote Sensing Letters*, 13 (11), pp. 1636 - 1640, 2016.
- [19] P. Tschandl et al., "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci Data*, vol. 5, 180161, 2018.
- [20] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.
- [21] R. Daneshjou, H. Barata, A. Betz-Stablein, et al., "Checklist for evaluation of image-based AI reports in dermatology: CLEAR Derm consensus guidelines," *Lancet Digital Health*, vol. 4, no. 6, pp. e442–e451, 2022.
- [22] A. S. Ahsan, M. A. Sifat, and M. Z. Islam, "Enhanced skin cancer diagnosis using optimized CNN architectures," *Comput Biol Med*, vol. 168, 107123, 2024.
- [23] H. Zhao, S. Li, Y. Yang, and J. Wu, "Accurate skin lesion classification using multimodal learning," *Front Oncol*, vol. 13, 1221067, 2023.
- [24] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.
- [25] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE Trans Med Imaging*, vol. 38, no. 9, pp. 2092–2103, 2019.