STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, October 2026, pp 3089–3102.
Published online in International Academic Press (www.IAPress.org)

Leakage-Controlled Evaluation of Handcrafted-Deep Feature Fusion and Confidence Stacking for Four-Class Skin Lesion Classification

Mohamed Amin Belal[1,*], Mohamed M. El-Gazzar[1], Ben Bella S. Tawfik[2], Sara. I. Ibrahim[2]

[1]*Department of Computer Science, Faculty of Computers and Artificial Intelligence, Modern University for Technology and Information (MTI), Cairo 4411601, Egypt*
[2]*Department of Computer Information Systems, Faculty of Computers and Informatics, Suez Canal University, Ismailia 41522, Egypt*

Abstract Automated classification of dermoscopic images remains difficult because benign and malignant lesions may share border, colour, shape, and texture characteristics. This study evaluates handcrafted features, frozen SqueezeNet embeddings, direct feature fusion, and sequential confidence stacking for basal cell carcinoma, melanoma, nevus, and pigmented benign keratosis. The dataset contained 1,445 images.
A leakage-controlled protocol used five outer folds and three inner folds; scaling, feature ranking, principal component analysis, class balancing, confidence generation, and adaptive cleaning were learned exclusively from outer-training data.
Direct fusion achieved the best pooled accuracy of 72.25% (95% CI: 69.97%-74.39%), with macro-F1 of 0.721, balanced accuracy of 0.716, cancer-versus-non-cancer ROC-AUC of 0.861, and PR-AUC of 0.870. Sequential stacking without cleaning achieved 68.03%, while adaptive cleaning reduced accuracy to 66.09%.
Direct fusion significantly outperformed the adaptive sequential model (exact McNemar $p = 3.89 \times 10^{-14}$; paired-fold $p = 5.18 \times 10^{-5}$). The end-to-end hierarchical model achieved 66.23% accuracy.
These corrected results show that direct handcrafted-deep fusion was more reliable than confidence stacking on this dataset and that automatic removal of low-margin training errors did not improve generalisation. External, patient-level validation remains necessary before clinical interpretation.

Keywords skin lesion classification; dermoscopy; hybrid features; SqueezeNet; confidence stacking; support vector machine; leakage-controlled cross-validation.

AMS 2010 subject classifications 62H30, 68T05, 92C40.

DOI: 10.19139/soic-2310-5070-4022

1. Introduction

Skin lesion assessment is visually demanding. Melanoma can resemble benign melanocytic lesions, while pigmented benign keratoses and some basal cell carcinomas may present overlapping colours, structures, or borders. Dermoscopy improves visual inspection, but diagnostic interpretation still depends on experience and may vary between readers. These characteristics have motivated computer-aided classification as a second-reading tool rather than a replacement for clinical judgement.

Recent CNN, transformer, and explainable systems have reported strong performance in dermoscopic image classification [1–3]. Their behaviour is influenced by dataset size, class distribution, acquisition conditions, and representation across skin tones [4, 5]. Current reporting guidance emphasises transparent split construction, uncertainty estimates, and reproducible model evaluation [6, 7]. A decision-support model should therefore be assessed with explicit controls for leakage, uncertainty, and failure patterns.

*Correspondence to: Mohamed Amin Belal (Email:abo_amin200@yahoo.com). Department of Computer Science, Faculty of Computers and Artificial Intelligence, Modern University for Technology and Information (MTI), Cairo 4411601, Egypt.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

Handcrafted descriptors remain relevant in moderate-size datasets because they encode visible gradient, texture, colour, and shape properties. Recent hybrid studies combine these descriptors with CNN embeddings, dimensionality reduction, and conventional classifiers [8–10]. Such components may be complementary, but additional decision-level confidence features are not necessarily informative when they are strongly correlated with the underlying deep representation.

The original submission did not clearly isolate learned preprocessing from final validation. The revised evaluation moves scaling, feature ranking, PCA, balancing, cleaning, and confidence generation inside each outer-training fold.

1.1. Motivation and contributions

The revised study asks whether confidence stacking improves direct handcrafted-deep fusion and whether confidence-based cleaning remains beneficial when restricted to training data. Its contributions are:

- A five-fold outer evaluation in which every learned operation is fitted only on the corresponding training fold.
- A controlled comparison of handcrafted-only, deep-only, direct-fusion, sequential no-cleaning, and sequential adaptive-cleaning variants.
- Out-of-fold confidence generation within each outer-training fold, preventing in-sample decision scores from entering meta-training.
- A transparent negative finding: direct fusion outperformed both confidence-stacking variants, and adaptive cleaning reduced rather than improved accuracy.
- Internal hierarchical evaluation, bootstrap confidence intervals, paired tests, uncertainty analysis, feature-selection stability, and high-resolution diagnostic figures.

2. Related Work and Positioning

2.1. Recent deep, transformer, and foundation-model approaches

Recent studies have evaluated vision transformers, handcrafted-deep fusion, explainable CNNs, hybrid attention, and comparative pretrained architectures for dermoscopic classification [1–3, 11, 12]. These works confirm that end-to-end deep learning is a strong baseline while also showing that performance depends on preprocessing, split construction, and dataset composition.

Research has expanded toward convolution-transformer mixtures and large pretrained visual encoders. Boundary-aware segmentation with convolutional and transformer classifiers has been studied [13], while PanDerm illustrates domain-specific multimodal foundation-model pretraining at substantially larger scale [14]. Such systems may improve transfer, but deployment still requires external validation and transparent reporting [15].

Bias and generalisation are central concerns. Recent work has analysed diagnostic disparities across skin tones and broader sources of imaging-AI bias [4, 5].

Newer datasets such as BCN20000 and SLICE-3D aim to broaden image diversity [16, 17], and uncertainty-aware skin-lesion evaluation has also been investigated [18].

Therefore, an internal four-class result should not be interpreted as evidence of clinical readiness.

2.2. Hybrid systems and SOIC studies

Recent handcrafted-deep fusion studies combine gradient, texture, colour, and shape measurements with CNN embeddings, PCA, and conventional classifiers [8–10]. The present evaluation differs by comparing simple direct concatenation with the reuse of inner out-of-fold confidence scores as decision-level features.

Relevant SOIC work includes SVM optimisation with CNN-derived features [10], fusion of deep image information and patient metadata [19], dedicated dermoscopic preprocessing [20], and transfer-learning comparison across imaging domains [21].

Recent segmentation and explainability studies also provide context for medical foundation models, joint segmentation-classification systems, and visual explanation methods [22, 24, 25].

The term meta-learning used in the original submission is replaced by confidence stacking or sequential feature fusion. The first-stage score vector is treated as a decision-level feature, not as conventional task-level meta-learning.

Table 1. Positioning relative to representative recent literature

Study	**Architecture / Feature Source**	**Data Basis**	**Relation to Present Evaluation**
Himel et al. [1]	Vision Transformer + segmentation	HAM10000 dermoscopy	Recent transformer baseline; different task
Naeem et al. [2]	Handcrafted + deep features	Dermoscopic datasets	Close feature-level comparator
Zahid et al. [8]	Classical descriptors + CNN + PCA/SVM	Skin-lesion benchmark	Direct hybrid fusion comparator
Silva et al. [10]	CNN features + dimensionality reduction + SVM	ISIC2017 and PH2	Relevant SOIC optimisation study
Omran et al. [19]	Deep image features + metadata	Clinical images and metadata	Hybrid multimodal SOIC study
Yan et al. [14]	Dermatology foundation model	¿2 million images, multiple institutions	Much larger pretraining scale
Present study	Handcrafted + frozen SqueezeNet + OOF scores	1,445 images; four classes	Same-fold ablation and leakage control

3. Dataset and Data Governance

The dataset contains 1,445 dermoscopic images organised in four class-labelled folders. It includes two cancer-related categories and two non-cancer categories. The exact class counts are shown in Table 2.

Table 2. Clinical Grouping of Skin Lesion Classes

Class	**Clinical Grouping**	**Images**
Basal cell carcinoma	Cancer	344
Melanoma	Cancer	407
Nevus	Non-cancer	309
Pigmented benign keratosis	Non-cancer	385
Total	–	1,445

Melanoma is the largest class and nevus is the smallest. Random oversampling is applied only to training data; outer-test images retain their original distribution.

The available files did not include patient identifiers, so the current results use image-level stratification.

Patient-grouped validation is recommended when identifiers become available.

3.1. Data integrity and split records

The revised computational package stores class counts, outer-fold assignments, dataset manifests, software information, selected features, cleaning logs, hyperparameters, and pooled out-of-fold predictions.

The random seed is fixed at 2026 to support repeatable split construction.

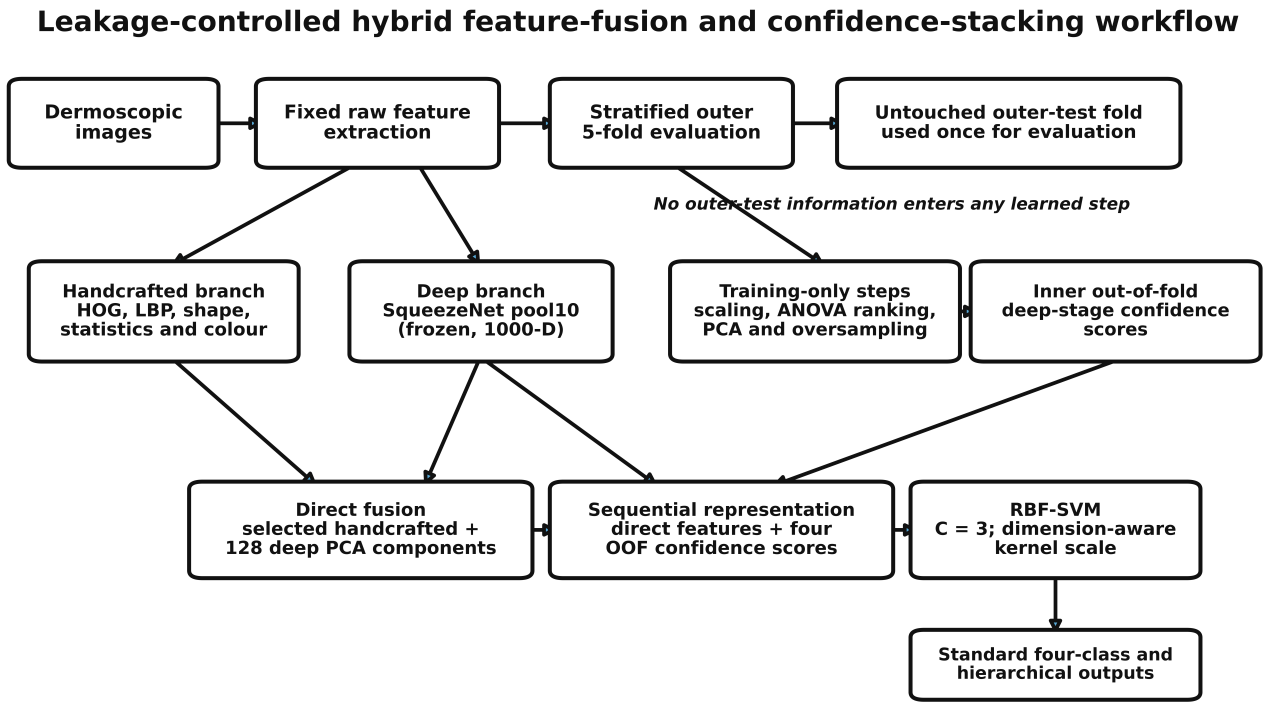

Figure 1. Leakage-controlled workflow used in the corrected experiment. Both direct fusion and confidence stacking are evaluated under the same outer folds

4. Method

4.1. Image preparation and automatic lesion mask

Each image is read in RGB form; grayscale images are replicated across three channels. The handcrafted branch uses a 96 x 96 grayscale image for HOG and a 128×128 image for LBP and intensity statistics. For shape and masked colour measurements, a reproducible 128×128 heuristic mask is generated using CIELAB luminance, global Otsu thresholding, adaptive dark-foreground thresholding, morphological opening and closing, hole filling, and small-component removal. The mask is not presented as a clinically validated segmentation model.

HOG and LBP are computed on resized grayscale images. The mask supports area, perimeter, circularity, compactness, eccentricity, solidity, extent, bounding-box measures, intensity statistics, entropy, grey-level histograms, Hu moments, RGB/HSV statistics, and HSV histograms.

4.2. Handcrafted representation

The raw handcrafted vector has 5,402 dimensions. It combines HOG, LBP, 14 shape descriptors, five intensity statistics, a 16-bin grayscale histogram, seven Hu moments, RGB/HSV means and standard deviations, and three 16-bin HSV histograms. Within each outer-training fold, variables are standardised and ranked using a univariate ANOVA F-score; the 250 highest-ranked variables are retained.

4.3. Frozen deep representation

The completed experiment used pretrained SqueezeNet, not ResNet-18. Images were resized to 227×227 pixels and embeddings were extracted from pool10. This yielded 1,000 frozen deep features per image. Within each outer-training fold, the deep variables were standardised and reduced to 128 principal components.

4.4. Leakage-controlled evaluation

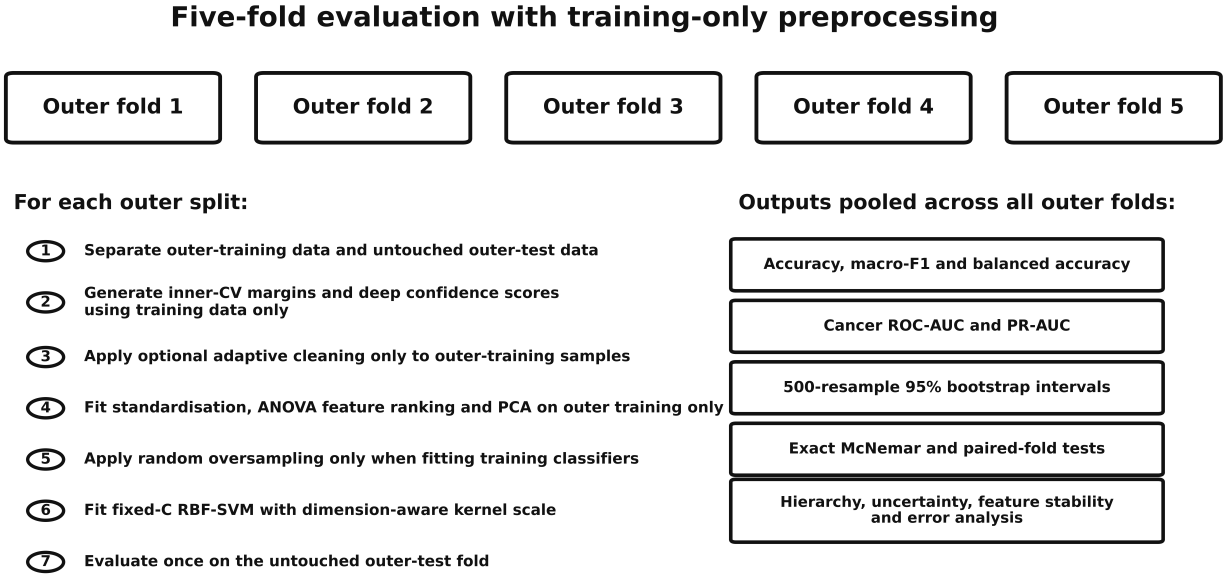

Figure 2. Five-fold outer evaluation. Every learned preprocessing step is fitted using outer-training data only

A stratified five-fold outer cross-validation estimates internal generalisation. For each split, the outer-test fold is separated before cleaning, standardisation, feature ranking, PCA, oversampling, and classifier fitting.

Three inner folds are used to generate training-only cleaning margins and out-of-fold deep-stage confidence scores. Predictions from all five outer-test folds are pooled for the principal estimates.

4.5. Adaptive confidence-based cleaning

The former fixed removal of 80 samples per class is discontinued. In each outer-training fold, a linear one-versus-all SVM produces inner out-of-fold predictions and true-class margins. A sample is eligible for removal only when it is misclassified and its margin is below the class-specific 10th percentile. Removal is capped at 10% of each training class and cannot reduce a class below 25 samples. The outer-test fold is never cleaned.

4.6. Compared feature configurations

Five configurations are evaluated: handcrafted only; deep only; direct fusion of 250 handcrafted and 128 deep components; sequential stacking without cleaning; and sequential stacking after adaptive cleaning. The sequential representation appends four inner out-of-fold deep-stage confidence scores and is reduced to at most 250 components before the final classifier.

All final classifiers use one-versus-one RBF-SVMs with box constraint $C = 3$. After training-fold standardisation, the Gaussian kernel scale is set to sigma = $\sqrt{\frac{d}{2}}$, where d is the fitted feature dimension.

This dimension-aware rule avoids the majority-class collapse produced by a fixed scale of one in high-dimensional feature spaces. Random oversampling is performed only on training data.

4.7. Hierarchical output

The secondary hierarchy first separates cancer (basal cell carcinoma or melanoma) from non-cancer (nevus or pigmented benign keratosis), then applies a branch-specific binary classifier. End-to-end accuracy includes routing errors. Branch accuracies computed using the true parent class are reported only as oracle diagnostic upper bounds.

4.8. Reproducibility settings

Table 3. Corrected implementation settings

Component	**Setting**
Random seed	2026
Outer / inner folds	5 / 3, stratified
Raw handcrafted dimension	5,402
Handcrafted selection	ANOVA F-score, top 250, training fold only
CNN embedding	SqueezeNet pool10, frozen, 1,000-D
Deep PCA	128 components, training fold only
Meta PCA	Up to 250 components, training fold only
Adaptive cleaning	10th-percentile margin; $\leq$10% per class; errors only
Final classifier	One-versus-one RBF-SVM
Box constraint	$C = 3$
Kernel scale	$\sigma = \sqrt{d/2}$ after standardisation
Balancing	Random oversampling of training data only
Bootstrap interval	500 resamples, percentile 95% CI
Execution environment	MATLAB R2022b; Windows 11; CPU; no GPU

4.9. Evaluation and statistical analysis

The reported metrics are pooled accuracy, macro-F1, balanced accuracy, per-class precision, recall and F1, cancer-versus-non-cancer ROC-AUC and PR-AUC, and a 500-resample percentile bootstrap interval for accuracy.

Direct fusion and adaptive sequential stacking are compared using the exact McNemar test on paired image predictions and a paired t-test across the five outer-fold accuracies. Confidence scores are normalised decision scores rather than calibrated posterior probabilities.

5. Results

5.1. Standard four-class performance and ablation

Direct fusion was the strongest configuration, with pooled accuracy of 72.25% (95% CI: 69.97%-74.39%). Its macro-F1 was 0.721 and balanced accuracy was 0.716.

The cancer-versus-non-cancer ROC-AUC and PR-AUC were 0.861 and 0.870, respectively. Both sequential configurations were less accurate than direct fusion.

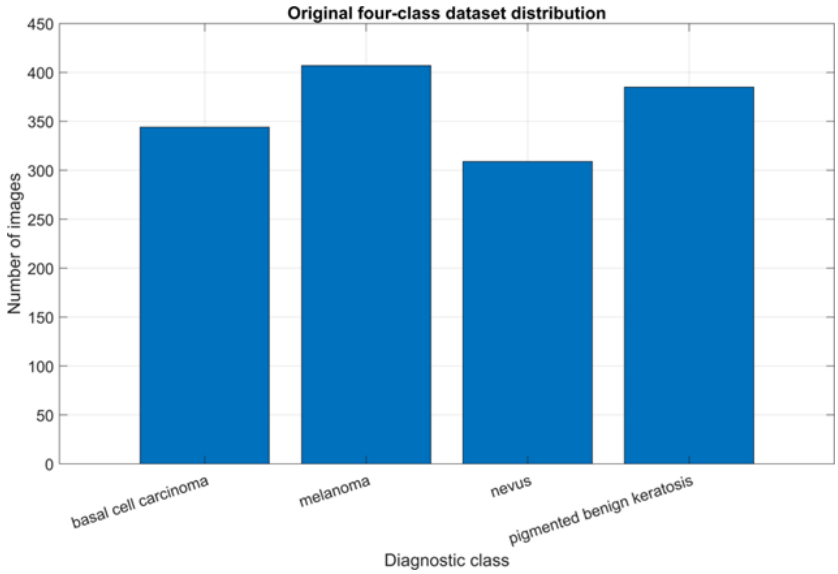

Figure 3. Original four-class dataset distribution. No outer-test sample is oversampled or removed

Table 4. Pooled out-of-fold performance under the corrected protocol

Variant	Accuracy (95% CI)	Macro-F1	Balanced Acc.	Cancer ROC-AUC	Cancer PR-AUC
Handcrafted only	68.10% (65.47%–70.45%)	0.680	0.676	0.813	0.815
Deep only	68.58% (66.30%–70.80%)	0.688	0.686	0.845	0.859
Direct fusion	72.25% (69.97%–74.39%)	0.721	0.716	0.861	0.870
Sequential, no cleaning	68.03% (65.81%–70.66%)	0.679	0.672	0.849	0.857
Sequential + adaptive cleaning	66.09% (63.53%–68.65%)	0.660	0.653	0.828	0.836

5.2. Paired statistical comparison

Compared with adaptive sequential stacking, direct fusion correctly classified 117 images that the sequential model missed, whereas the sequential model corrected 28 direct-fusion errors. The exact McNemar test and paired-fold test both favoured direct fusion.

Table 5. Direct fusion versus adaptive sequential stacking

Statistic	Value
Direct wrong / sequential correct (b)	28
Direct correct / sequential wrong (c)	117
Exact McNemar p-value	3.89×10^{-14}
Paired-fold t statistic (df)	$-18.36\,(4)$
Paired-fold p-value	5.18×10^{-5}

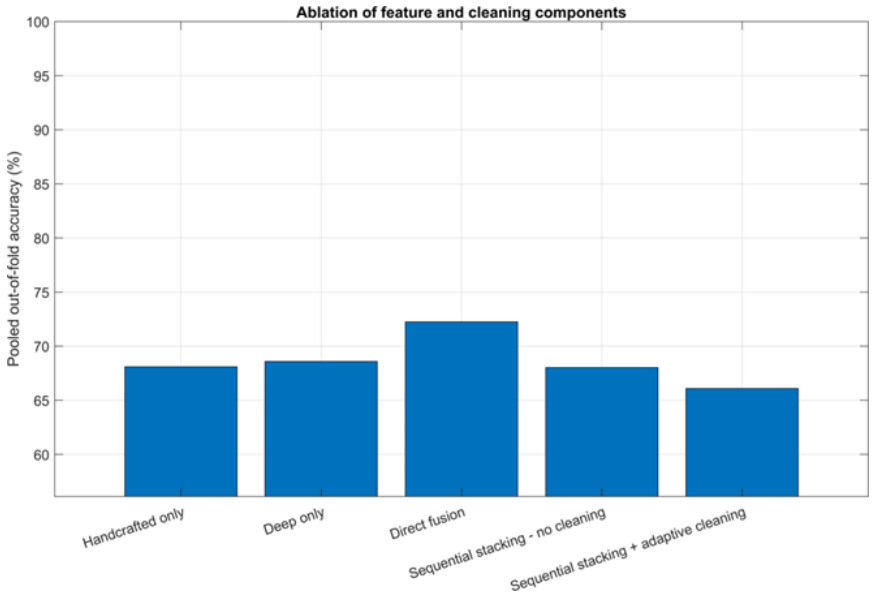

Figure 4. Ablation results. Direct fusion outperformed both confidence-stacking configurations

5.3. Effect of adaptive cleaning

Adaptive cleaning removed 113 training samples in every outer fold: approximately 27 basal cell carcinoma, 32 melanoma, 24 nevus, and 30 pigmented benign keratosis samples.

Accuracy decreased from 68.03% without cleaning to 66.09% with cleaning, a reduction of 1.94 percentage points. Thus, the training-only cleaning rule did not add value in the corrected experiment.

Table 6. Cleaning ablation and average removal counts

Condition	Accuracy	Mean Removed per Fold	Change
Sequential stacking without cleaning	68.03%	0	–
Sequential stacking with adaptive cleaning	66.09%	113	-1.94 percentage points
Average removed: basal cell carcinoma	–	27	–
Average removed: melanoma	–	32	–
Average removed: nevus	–	24	–
Average removed: pigmented benign keratosis	–	30	–

5.4. Hierarchical performance

The end-to-end hierarchy achieved 66.23% accuracy (95% CI: 63.74%-68.65%), below direct fusion. The Level-1 cancer/non-cancer mean accuracy was 77.09% +/- 2.08%. Oracle branch accuracies were 86.29% for the cancer branch and 80.55% for the non-cancer branch; these values assume correct routing and are not deployment estimates.

5.5. Per-class direct-fusion results

Direct fusion produced its highest recall for melanoma (82.1%) and its lowest recall for nevus (60.2%). This pattern indicates that the four-class task remained difficult even though the cancer-versus-non-cancer ranking was comparatively stronger.

Table 7. Hierarchical evaluation

Measure	Result
End-to-end four-class accuracy	66.23% (63.74%–68.65%)
Fold mean ± SD	66.23% ± 1.31%
Level-1 cancer/non-cancer accuracy	77.09% ± 2.08%
Cancer branch oracle accuracy	86.29%
Non-cancer branch oracle accuracy	80.55%

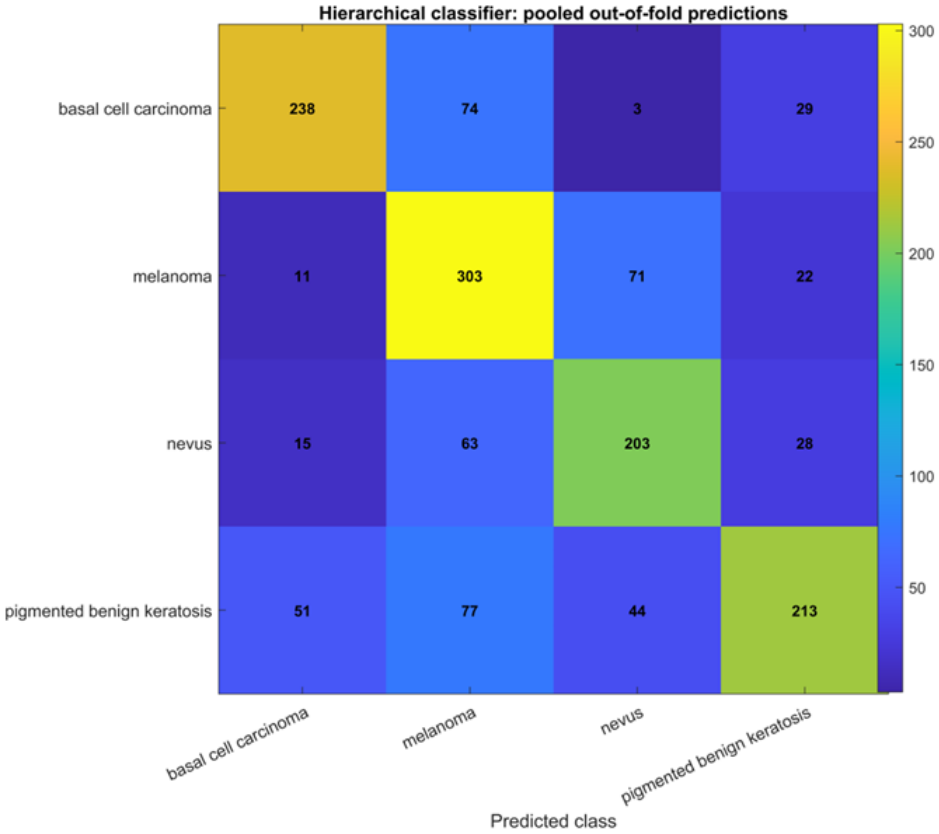

Figure 5. Pooled end-to-end hierarchical confusion matrix. Routing errors are included

Table 8. Per-class metrics for the best-performing direct-fusion model

Class	Support	Precision	Recall	F1
Basal cell carcinoma	344	0.800	0.756	0.777
Melanoma	407	0.654	0.821	0.728
Nevus	309	0.769	0.602	0.675
Pigmented benign keratosis	385	0.719	0.686	0.702

5.6. Uncertainty, feature stability, and error patterns

For the adaptive sequential model, the selective risk at 80% coverage was 24.83%, corresponding to 75.17% accuracy among the 80% most confident predictions. The most frequent errors were pigmented benign keratosis classified as melanoma (99 images), nevus classified as melanoma (96), and basal cell carcinoma classified as melanoma (76). These findings show a tendency toward melanoma predictions in ambiguous cases.

Table 9. Uncertainty and implementation summary

Item	**Result**
Selective risk at 80% coverage	24.83%
Accuracy among most confident 80%	75.17%
Most frequent confusion	PBK → melanoma: 99
Second most frequent confusion	Nevus → melanoma: 96
Software	MATLAB R2022b on Windows 11
Execution	CPU; no GPU selected
Feature families selected consistently	HOG, LBP, intensity and colour descriptors

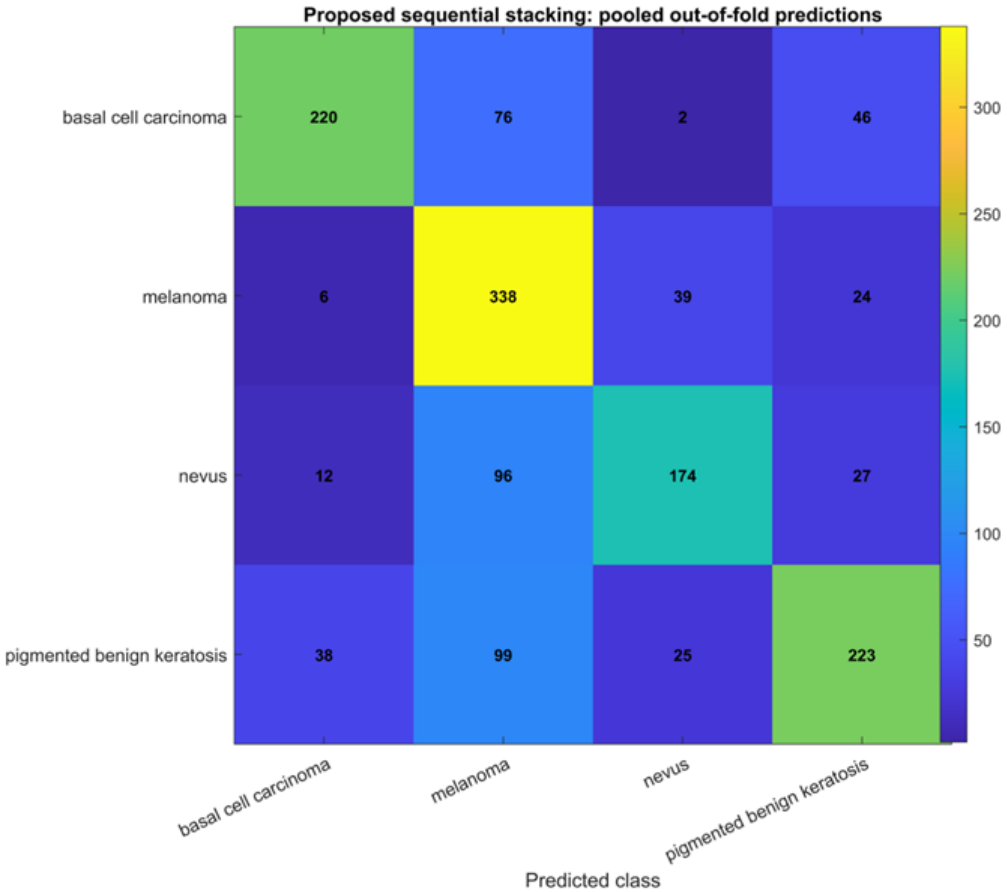

Figure 6. Confusion matrix for the originally proposed sequential model with adaptive cleaning. The model showed a strong melanoma prediction bias in ambiguous cases

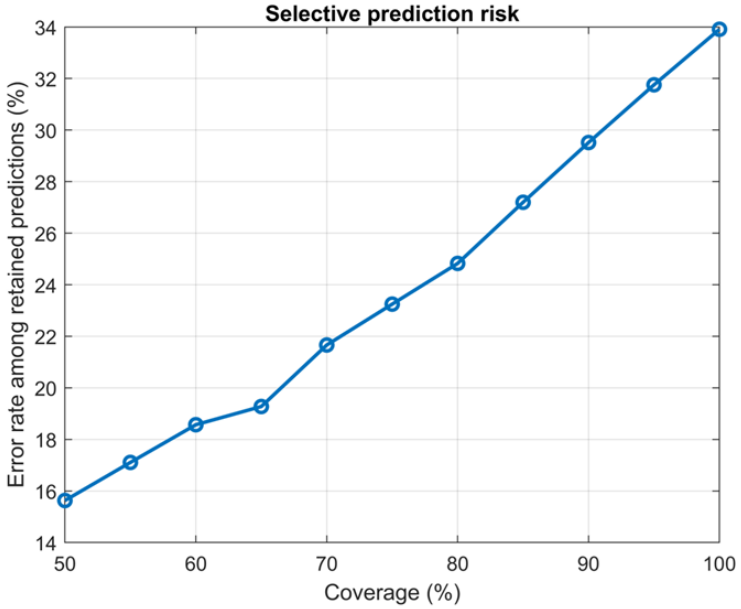

Figure 7. Selective-risk curve for the adaptive sequential model. Withholding low-confidence cases reduced error among retained cases

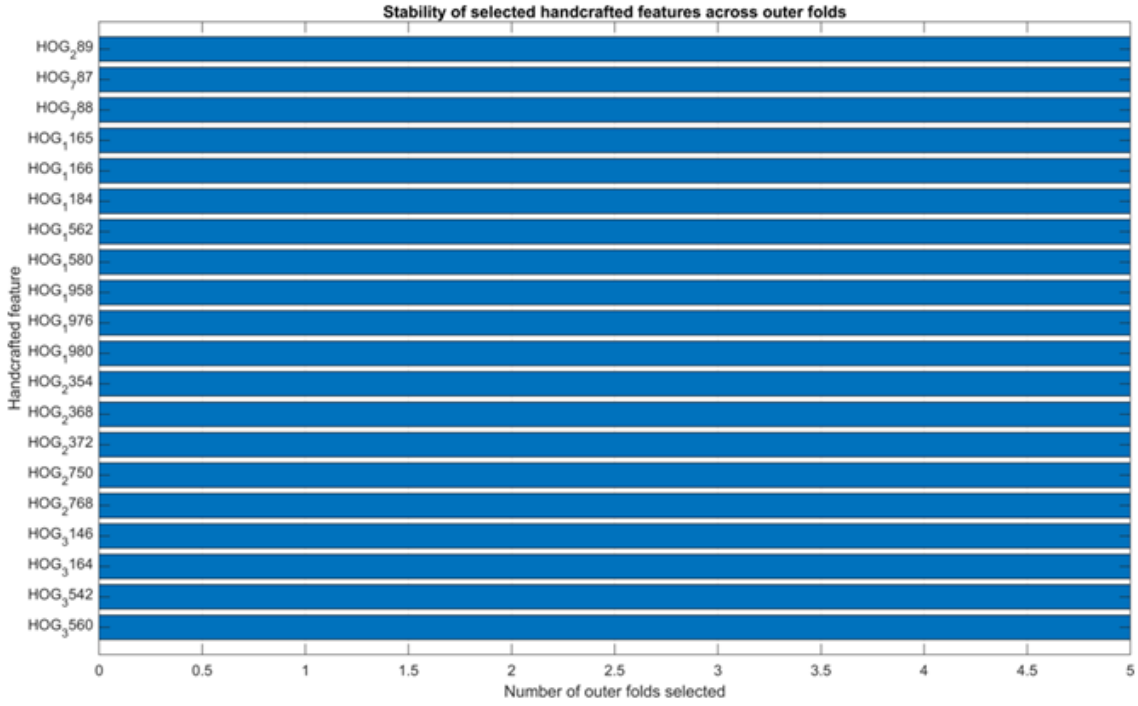

Figure 8. Stability of the most frequently selected handcrafted variables across the five outer folds

6. Discussion

The corrected experiment does not support the original hypothesis that sequential confidence stacking improves direct fusion.

Direct fusion achieved 72.25% accuracy, whereas stacking without cleaning achieved 68.03% and adaptive cleaning achieved 66.09%.

The paired tests show that this difference is not random variation in favour of the proposed sequential model; it is a statistically significant advantage for direct fusion.

A plausible explanation is that the four confidence scores were derived from the same deep representation already present in the fused vector.

They may therefore add redundant or noisy information rather than independent evidence. The extra PCA stage required by the sequential representation may also compress useful handcrafted-deep structure.

The adaptive cleaning result is also cautionary. Removing approximately 9.8% of each outer-training set did not improve generalisation. Low-margin errors may include genuine difficult phenotypes rather than mislabeled observations.

In a clinical research workflow, such samples should be reviewed by qualified annotators instead of being removed automatically.

The hierarchy did not improve the four-class endpoint because routing errors propagated to the subtype branches. The oracle branch results are substantially higher, confirming that part of the loss occurred at Level 1.

The hierarchy may still be useful in a triage-oriented study, but this dataset does not show an end-to-end accuracy advantage.

The direct-fusion model demonstrates complementary value between handcrafted and frozen deep features: it exceeded both single-branch models by approximately four percentage points.

Nevertheless, the corrected 72.25% accuracy is far below the withdrawn values from the leakage-prone programs. This difference illustrates why strict separation of preprocessing and evaluation is essential in moderate-size medical-imaging studies.

7. Limitations and Clinical Relevance

The study provides internal image-level validation only. Dataset provenance, patient identifiers, acquisition devices, and demographic or skin-tone subgroup information were not available in the supplied files.

If multiple images originate from the same patient, image-level folds may be optimistic. Independent external and patient-grouped validation is therefore required.

SqueezeNet was used as a frozen feature extractor because the ResNet-18 support package was unavailable during the completed run. Results should not be generalized to deeper or dermatology-specific encoders.

The core run also did not include an end-to-end fine-tuned CNN comparator, Grad-CAM analysis, fixed 0/20/40/80 cleaning sensitivity, or an end-to-end inference-time benchmark. These analyses are not claimed in the revised manuscript and remain future work.

The automatic mask is a reproducible preprocessing heuristic rather than a clinically validated segmentation system. Normalised SVM scores are useful for ranking confidence but are not calibrated probabilities.

The present work is a computational evaluation and should not be used to exclude biopsy, specialist assessment, or follow-up.

8. Reproducibility and Availability

The completed experiment was generated with Model_8_4_2026_CORE_FAST_V4.m using MATLAB R2022b on Windows 11 with CPU execution. The output package includes fold assignments, dataset manifests, class counts, cleaning logs, selected variables, model settings, pooled metrics, paired tests, uncertainty records, misclassified cases, software information, and 600-dpi figures.

The exact dataset source, licence, and code repository address should be inserted by the authors in the journal template before submission.

9. Conclusion

Under leakage-controlled five-fold evaluation, direct handcrafted-deep fusion was the best-performing configuration, with 72.25% accuracy, macro-F1 of 0.721, balanced accuracy of 0.716, cancer ROC-AUC of 0.861, and PR-AUC of 0.870.

Sequential confidence stacking did not improve direct fusion, and adaptive cleaning further reduced accuracy. The end-to-end hierarchy also remained below direct fusion.

The main contribution of the revision is therefore a transparent component evaluation that replaces optimistic leakage-prone estimates with reproducible evidence. External, patient-level validation and stronger end-to-end CNN baselines are required before clinical claims can be considered.

REFERENCES

1. G. M. S. Himel, M. M. Islam, K. A. Al-Aff, S. I. Karim, and M. K. U. Sikder, *Skin cancer segmentation and classification using vision transformer for automatic analysis in dermatoscopy-based noninvasive digital system*, International Journal of Biomedical Imaging, vol. 2024, Art. no. 3022192, 2024. doi:10.1155/2024/3022192.
2. A. Naeem, T. Anees, M. Khalil, K. Zahra, R. A. Naqvi, and S.-W. Lee, *SNC_Net: Skin cancer detection by integrating handcrafted and deep learning-based features using dermoscopy images*, Mathematics, vol. 12, no. 7, Art. no. 1030, 2024. doi:10.3390/math12071030.
3. S. A. H. Shah et al., *Explainable AI-based skin cancer detection using CNN, particle swarm optimization and machine learning*, Journal of Imaging, vol. 10, no. 12, Art. no. 332, 2024. doi:10.3390/jimaging10120332.
4. M. Groh et al., *Deep learning-aided decision support for diagnosis of skin disease across skin tones*, Nature Medicine, vol. 30, pp. 573–583, 2024. doi:10.1038/s41591-023-02728-3.
5. B. Koçak et al., *Bias in artificial intelligence for medical imaging: Fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects*, Diagnostic and Interventional Radiology, vol. 31, no. 2, pp. 75–88, 2025. doi:10.4274/dir.2024.242854.
6. G. S. Collins et al., *TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods*, BMJ, vol. 385, Art. no. e078378, 2024. doi:10.1136/bmj-2023-078378.
7. A. S. Tejani et al., *Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update*, Radiology: Artificial Intelligence, vol. 6, no. 4, Art. no. e240300, 2024. doi:10.1148/ryai.240300.
8. M. Zahid, M. Rziza, and R. Alaoui, *Skin lesion classification using hybrid feature extraction based on classical and deep learning methods*, BioMedInformatics, vol. 5, no. 3, Art. no. 41, 2025. doi:10.3390/biomedinformatics5030041.
9. K. Behara, E. Bhero, and J. T. Agee, *An improved skin lesion classification using a hybrid approach with active contour snake model and lightweight attention-guided capsule networks*, Diagnostics, vol. 14, no. 6, Art. no. 636, 2024. doi:10.3390/diagnostics14060636.
10. J. Silva, A. Alves, P. Santos, and L. Matioli, *A new SVM solver applied to skin lesion classification*, Statistics, Optimization & Information Computing, vol. 12, no. 4, pp. 1149–1172, 2024. doi:10.19139/soic-2310-5070-2005.
11. H. T. Halawani et al., *Enhanced early skin cancer detection through fusion of vision transformer and CNN features using hybrid attention of EViT-Dens169*, Scientific Reports, vol. 15, Art. no. 34776, 2025. doi:10.1038/s41598-025-18570-1.
12. A. Alrabai, A. Echtioui, and F. Kallel, *Exploring pre-trained models for skin cancer classification*, Applied System Innovation, vol. 8, no. 2, Art. no. 35, 2025. doi:10.3390/asi8020035.
13. J. Amin et al., *Skin-lesion segmentation using boundary-aware segmentation network and classification based on a mixture of convolutional and transformer neural networks*, Frontiers in Medicine, vol. 12, Art. no. 1524146, 2025. doi:10.3389/fmed.2025.1524146.
14. S. Yan et al., *A multimodal vision foundation model for clinical dermatology*, Nature Medicine, vol. 31, pp. 2691–2702, 2025. doi:10.1038/s41591-025-03747-y.
15. H. Gui, J. A. Omiye, C. T. Chang, and R. Daneshjou, *The promises and perils of foundation models in dermatology*, Journal of Investigative Dermatology, vol. 144, no. 7, pp. 1440–1448, 2024. doi:10.1016/j.jid.2023.12.019.
16. C. Hernández-Pérez et al., *BCN20000: Dermoscopic lesions in the wild*, Scientific Data, vol. 11, Art. no. 641, 2024. doi:10.1038/s41597-024-03387-w.
17. N. R. Kurtansky et al., *The SLICE-3D dataset: 400,000 skin lesion image crops extracted from 3D total body photography for skin cancer detection*, Scientific Data, vol. 11, Art. no. 884, 2024. doi:10.1038/s41597-024-03743-w.
18. J. Fayyad, S. Alijani, and H. Najjaran, *Empirical validation of conformal prediction for trustworthy skin lesions classification*, Computer Methods and Programs in Biomedicine, vol. 253, Art. no. 108231, 2024. doi:10.1016/j.cmpb.2024.108231.
19. A. S. A. E. A. Omran, M. M. El-Gazzar, and M. M. Saeid, *A hybrid fusion approach for skin cancer detection using deep learning on clinical images and machine learning on patient metadata*, Statistics, Optimization & Information Computing, vol. 14, no. 4, pp. 2091–2103, 2025. doi:10.19139/soic-2310-5070-2811.

20. Y. El Idrissi El-Bouzaidi, S. El Khamlichi, and O. Abdoun, *Diagnosis with deep learning and a novel pre-processing strategy on a large-scale dermoscopic image dataset*, Statistics, Optimization & Information Computing, vol. 15, no. 3, pp. 2069–2085, 2026. doi:10.19139/soic-2310-5070-3122.
21. F. F. Alkhalid and A. M. Hasan, *The effect of applying transfer learning approach on different domains: Medical and non-medical imaging (skin cancer and flower types)*, Statistics, Optimization & Information Computing, vol. 14, no. 5, pp. 2571–2589, 2025. doi:10.19139/soic-2310-5070-2631.
22. J. Ma et al., *Segment Anything in Medical Images*, Nature Communications, vol. 15, Art. no. 654, 2024. doi:10.1038/s41467-024-44824-z.
23. M. Fiaz et al., *An explainable hybrid deep learning framework for precise skin lesion segmentation and multi-class classification*, Frontiers in Medicine, vol. 12, Art. no. 1681542, 2025. doi:10.3389/fmed.2025.1681542.
24. S. M. S. I. Badhon et al., *Explainable AI for skin disease classification using gradient-weighted class activation mapping and transfer learning in digital health to identify contours*, Digital Health, vol. 11, Art. no. 20552076251404523, 2025. doi:10.1177/20552076251404523.
25. A. S. Al-Waisy et al., *A deep learning framework for automated early diagnosis and classification of skin cancer lesions in dermoscopy images*, Scientific Reports, vol. 15, Art. no. 31234, 2025. doi:10.1038/s41598-025-15655-9.