

MedLite XAI A Metadata Aware Calibrated and Explainable Lightweight Framework for Multiclass Skin Lesion Classification

Baibhav Tiwari¹ Prakash Kumar Maravi² Kamlesh Raghuvanshi³ Surabhi Karsoliya⁴

¹M Tech Research Scholar ²Assistant Professor Department of AIML ³Head of Department ⁴Management Member

Technocrats Institute of Technology-Excellence Bhopal Madhya Pradesh India

Abstract: Automated dermoscopic classification can support skin lesion assessment while reliable research requires leakage safe partitions class sensitive evaluation calibrated confidence interpretable predictions and practical computational cost; therefore this work presents MedLite XAI which combines a MobileNetV3 Small image encoder with patient metadata for seven class classification on HAM10000; a deterministic lesion grouped split allocated 7012 training images 1502 validation images and 1501 locked test images while model selection used validation macro F1 only; the selected metadata aware cross entropy model obtained 74.65 percent validation macro F1 and achieved 84.08 percent accuracy 71.59 percent macro F1 and 71.34 percent balanced accuracy on the final test set; temperature scaling reduced expected calibration error from 6.69 percent to 2.09 percent while preserving class predictions and Grad CAM analysis across 20 validation cases produced mean horizontal flip stability of 0.969; consequently MedLite XAI provides a reproducible lightweight research framework although external and prospective clinical validation remain necessary.

Index Terms: skin lesion classification; dermoscopy; HAM10000; lightweight deep learning; metadata fusion; calibration; explainable artificial intelligence.

I INTRODUCTION

Skin cancer assessment is a high impact clinical problem while dermoscopy provides visual information that can be difficult to interpret consistently; deep neural networks have demonstrated strong performance in dermatology benchmarks although accuracy alone does not establish reliable clinical behavior [1] [2].

HAM10000 contains repeated images for some lesions with strong class imbalance and limited patient metadata; consequently an image level random split can create information leakage while aggregate accuracy can conceal weak recognition of uncommon classes [3].

Lightweight encoders can reduce computational cost while patient age sex and anatomical location can add contextual information to the image representation; however the contribution of metadata must be evaluated through controlled ablation and every preprocessing operation must be fitted on training data only [4] [5].

Confidence reliability and visual explanation are also important because a model can provide accurate labels with misleading certainty while a plausible heatmap does not independently establish clinical validity; therefore calibration and explanation stability should complement discrimination metrics [6] [7].

This study contributes a lesion grouped and test locked protocol a compact image and metadata architecture controlled loss ablations validation only temperature scaling quantitative explanation stability and one final frozen test evaluation.

II RELATED WORK

Convolutional transfer learning has established strong baselines for dermoscopic analysis while MobileNet and EfficientNet architectures offer practical accuracy and efficiency tradeoffs; recent studies also use attention transformers feature fusion ensembles and knowledge distillation [8] [9] [10].

Class imbalance is addressed through augmentation resampling class weighting focal objectives and feature fusion while direct comparison remains difficult because studies use different datasets class mappings split units and metrics [11] [12].

Metadata fusion and explainable frameworks increasingly combine visual features with patient information while calibration and lesion grouped split governance remain less consistently reported; MedLite XAI therefore evaluates these requirements within one controlled experiment [13] [14].

III MATERIALS AND METHODS

A Dataset and Data Audit

HAM10000 contains 10015 dermoscopic images from 7470 lesions across seven diagnostic categories consisting of actinic keratoses and intraepithelial carcinoma basal cell carcinoma benign keratosis like lesions dermatofibroma melanoma melanocytic nevi and vascular lesions [3].

The local audit identified 57 missing age records 57 unknown sex records 234 unknown localization records and two exact duplicate content pairs while both duplicate pairs remained within the same lesion and partition.

B Leakage Safe Split

A deterministic lesion grouped stratified split with seed 20260727 created 7012 training images from 5224 lesions 1502 validation images from 1122 lesions and 1501 test images from 1124 lesions; the final test set was excluded from preprocessing model selection calibration fitting threshold selection and architecture decisions.

C Preprocessing and Metadata Encoding

Images were resized and normalized for the selected encoder while training only augmentation supported generalization; age missing age status sex and anatomical localization were encoded through a compact metadata network with unknown values mapped to explicit states.

Augmentation was restricted to the training partition so that validation and test observations retained their original

evaluation role; furthermore metadata statistics and categorical mappings were learned from training records only which prevented information from later partitions influencing feature preparation.

D Dataset Governance and Reproducibility

Every image was linked to its lesion identifier before partitioning while duplicate content and missing metadata were documented in machine readable audit artifacts; consequently the same manifests can be reused across models and paired comparisons without changing the evaluation population.

TABLE I LESION GROUPED DATASET PARTITIONS

Partition	Images	Unique lesions	Intended use
Training	7012	5224	Model fitting and preprocessing
Validation	1502	1122	Selection calibration and analysis
Test	1501	1124	One final frozen evaluation
Total	10015	7470	Complete HAM10000 collection

E Models and Proposed Architecture

The experimental comparison included a small convolutional network MobileNetV3 Small EfficientNet B0 an image only MedLite variant and metadata aware MedLite variants; the proposed model concatenates compact image features with metadata features before the final seven class classifier.

MobileNetV3 Small provides the image branch because depthwise operations and compact inverted residual blocks reduce parameter demand while retaining useful visual representations; the metadata branch transforms the structured variables through a small multilayer network before fusion.

The fused representation is passed to the final classifier while dropout and regularized optimization reduce overfitting; this late fusion design keeps the image and metadata pathways interpretable enough for controlled removal of the metadata contribution.

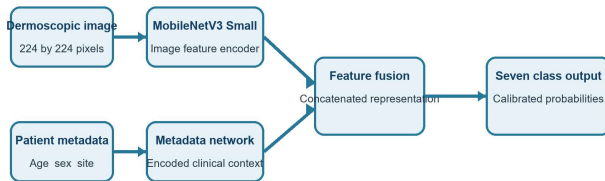


Fig 1 MedLite XAI architecture with image and metadata fusion

F Training and Model Selection

AdamW optimization early stopping and a validation driven learning rate schedule were applied consistently while macro F1 served as the primary selection metric; cross entropy weighted cross entropy and class balanced focal objectives were compared without using the locked test partition.

The weighted objective increased the contribution of minority classes while the class balanced focal objective reduced emphasis on easy frequent examples; nevertheless the

final loss was selected from validation macro F1 rather than theoretical preference or test performance.

Checkpoint selection used the best validation macro F1 within each run while incomplete continuation experiments were retained only as supplementary evidence; this rule prevented interrupted optimization histories from influencing the primary comparison.

G Evaluation Metrics

Accuracy measured total correctness while balanced accuracy averaged class recall and macro F1 averaged class specific harmonic means; these class sensitive measures were emphasized because melanocytic nevi dominate the dataset and can otherwise control the aggregate result.

Macro precision weighted F1 confusion matrices parameter count checkpoint size and single image latency provided supporting evidence; no single metric was treated as sufficient for model selection or deployment interpretation.

H Calibration Explainability and Statistics

A scalar temperature was fitted to validation logits while expected calibration error with 15 bins negative log likelihood and multiclass Brier score measured probability reliability; Grad CAM and Grad CAM plus plus were generated for 20 validation images while horizontal flip stability assessed explanation consistency.

Lesion clustered bootstrap confidence intervals used 2000 replicates while paired fixed sample comparisons and exact McNemar tests were calculated against the small convolutional reference with Holm adjustment.

Bootstrap resampling operated at lesion level so repeated images from the same lesion moved together within each replicate; this approach aligned uncertainty estimation with the grouped structure used during partitioning.

I Implementation Environment

The experiments were executed through configuration driven Python scripts with saved run metadata histories predictions and checkpoints; final result generation consumed the frozen artifacts so reported tables and figures did not require manual transcription.

IV RESULTS

A Validation Model Selection

The metadata aware MedLite cross entropy model achieved the strongest validation macro F1 of 74.65 percent with a lesion clustered 95 percent confidence interval from 69.10 percent to 78.63 percent; EfficientNet B0 achieved 73.65 percent while MobileNetV3 Small achieved 72.07 percent.

TABLE II VALIDATION MODEL COMPARISON

Model	Accuracy	Macro F1	Parameters	Latency
Small CNN	74.10	45.53	94599	2.39 ms
MobileNetV3 Small	83.02	72.07	1525031	20.13 ms
EfficientNet B0	83.16	73.65	4016515	29.72 ms
MedLite image only	82.56	70.21	1003031	11.27 ms
MedLite metadata CE	84.55	74.65	1005079	26.04 ms
MedLite weighted CE	81.76	72.30	1005079	38.50 ms

Validation Macro-F1 Across Full Models

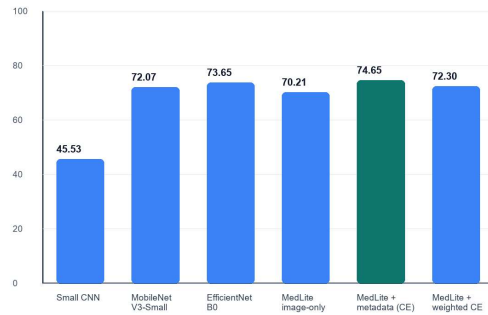


Fig 2 Validation comparison of baseline and MedLite variants

B Statistical Comparison

The selected metadata aware cross entropy model improved macro F1 by 29.11 percentage points relative to the small convolutional reference while the lesion clustered 95 percent interval for the improvement extended from 23.57 to 34.30 percentage points; the Holm adjusted McNemar probability was 2.96 multiplied by ten to the power minus 19.

The interval and paired test support a clear improvement over the weakest baseline while overlapping intervals among stronger transfer learning models require a more restrained interpretation; consequently the selected model is described as best within this validation protocol rather than universally superior.

C Final Locked Test Performance

The selected model achieved 84.08 percent accuracy 71.59 percent macro F1 71.34 percent balanced accuracy 72.51 percent macro precision and 83.62 percent weighted F1 on 1501 test images from 1124 lesions.

The largest observed confusion occurred between melanoma and melanocytic nevus while minority class errors remained clinically important despite strong aggregate accuracy; therefore the confusion matrix and class sensitive metrics are interpreted together.

The difference between validation macro F1 and final test macro F1 was reported transparently because the test set was never used to revise the architecture loss preprocessing or calibration strategy; this reduction illustrates why validation selection and final performance must remain separate.

Final Test Confusion Matrix (Calibrated Model)

	akiec	bcc	bkl	df	mel	nv	vasc
akiec	33	3	8	1	2	2	0
bcc	3	57	2	2	1	12	0
bkl	12	7	105	1	21	19	0
df	2	2	0	9	0	4	0
mel	9	4	10	0	86	58	0
nv	1	5	14	1	29	952	3
vasc	0	0	0	0	0	1	20

Rows: true class; columns: predicted class.

Fig 3 Confusion matrix from the final locked test evaluation

D Calibration

Validation fitted temperature scaling used a temperature of 1.568 while test expected calibration error decreased from 6.69 percent to 2.09 percent negative log likelihood decreased from 0.522 to 0.456 and Brier score decreased from 0.241 to 0.232 without changing predicted classes.

TABLE III FINAL TEST AND CALIBRATION RESULTS

Metric	Raw	Calibrated
Accuracy percent	84.08	84.08
Macro F1 percent	71.59	71.59
Balanced accuracy percent	71.34	71.34
ECE percent	6.69	2.09
NLL	0.522	0.456
Brier score	0.241	0.232

Final Test Calibration: Before and After Temperature Scaling



Fig 4 Reliability before and after validation fitted temperature scaling

E Explainability

Grad CAM and Grad CAM plus plus overlays were generated for 20 validation cases while mean horizontal flip stability reached 0.969; these maps provide model inspection evidence although they do not establish causal reasoning or clinical validation.

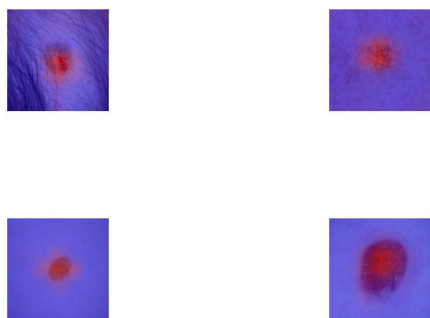


Fig 5 Representative Grad CAM overlays from the validation analysis

F Efficiency and Deployment Evidence

The selected model contained 1005079 trainable parameters which was smaller than MobileNetV3 Small and approximately one quarter of EfficientNet B0; measured latency was higher than the image only MedLite variant because metadata processing and the recorded runtime path added overhead.

The result demonstrates a practical parameter advantage while latency values remain hardware and software dependent; therefore deployment claims should be based on measurements from the intended device rather than parameter count alone.

G Error Analysis

Error review identified confusion among visually related pigmented categories while rare class estimates remained less stable because fewer lesions contributed to each measure; consequently future evaluation should include external data lesion masks and subgroup evidence.

Incorrect high confidence predictions are particularly important because they combine discrimination and reliability failure; calibrated probabilities reduced aggregate confidence error while case level review remains necessary for safety focused research.

V DISCUSSION

The results support the central claim that compact image and metadata fusion can improve class sensitive validation performance while retaining a modest parameter count; nevertheless the gain must be interpreted within the fixed HAM10000 protocol rather than as universal superiority.

Metadata may improve predictions by representing age sex and anatomical context while it can also encode dataset specific prevalence patterns; therefore external validation subgroup analysis and controlled missingness experiments remain important.

Temperature scaling substantially improved probability reliability while preserving discrimination which supports separate evaluation of calibration and classification; however calibration can shift across populations and acquisition settings.

Explanation stability was high under horizontal reflection while stability does not prove clinical relevance; future work

should compare heatmaps with lesion masks expert review counterfactual tests and failure focused sampling.

The study remains limited by retrospective public data absence of external testing incomplete demographic coverage and lack of prospective clinician evaluation; additionally the framework is a research prototype and must not be used for autonomous diagnosis.

A Comparison with Recent Literature

Recent fusion attention and transformer studies report strong results while direct numerical ranking is inappropriate when split policies and test populations differ; the main contribution of this work is the combined protocol for grouping calibration explanation stability efficiency and locked test governance rather than a claim based on accuracy alone.

B Clinical and Research Implications

Reliable probabilities can support threshold based review and uncertainty communication while class sensitive metrics expose minority category behavior; nevertheless any clinical workflow would require external validation prospective assessment clinician oversight and analysis of consequences at specific decision thresholds.

C Threats to Validity

Internal validity is strengthened by fixed manifests and test isolation while external validity is limited by one retrospective dataset; construct validity is limited because diagnostic labels and metadata may not represent the complete clinical decision process.

Reproducibility is supported by saved configurations predictions and statistical artifacts while exact latency can vary with hardware libraries and background processes; explanation stability under one transformation also cannot represent every form of robustness.

D Future Work

Future work should test the frozen model on independent dermoscopic collections and should evaluate subgroup behavior across age sex site and skin tone where reliable metadata are available; calibration should be reassessed after every population or acquisition shift.

Prospective research should compare image only metadata aware and clinician assisted workflows while explanation evaluation should incorporate lesion masks expert scoring perturbation tests and documented failure cases.

VI CONCLUSION

MedLite XAI combines leakage safe evaluation lightweight image encoding patient metadata calibration and explainability within one reproducible multiclass skin lesion framework; the selected model achieved 84.08 percent test accuracy 71.59 percent macro F1 and calibrated expected calibration error of 2.09 percent while external and prospective validation remain required before clinical translation.

REFERENCES

- [1] A Esteva et al. Dermatologist level classification of skin cancer with deep neural networks Nature vol 542 pp 115 to 118 2017

- [2] H A Haenssle et al Man against machine diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to dermatologists *Annals of Oncology* 2018
- [3] P Tschandl C Rosendahl and H Kittler The HAM10000 dataset a large collection of multisource dermoscopic images of common pigmented skin lesions *Scientific Data* 2018
- [4] A Howard et al Searching for MobileNetV3 *Proceedings of the IEEE International Conference on Computer Vision* 2019
- [5] M Tan and Q V Le EfficientNet rethinking model scaling for convolutional neural networks *Proceedings of the International Conference on Machine Learning* 2019
- [6] R R Selvaraju et al Grad CAM visual explanations from deep networks via gradient based localization *Proceedings of the IEEE International Conference on Computer Vision* 2017
- [7] C Guo G Pleiss Y Sun and K Q Weinberger On calibration of modern neural networks *Proceedings of the International Conference on Machine Learning* 2017
- [8] Y Nie et al A deep CNN transformer hybrid model for skin lesion classification of dermoscopic images using focal loss 2022
- [9] H Cheng J Lian and W Jiao Enhanced MobileNet for skin cancer image classification with fused spatial channel attention mechanism 2024
- [10] D P Yadav et al Dual scale lightweight cross attention transformer for skin lesion classification 2024
- [11] Z Hu W Mei H Chen and W Hou Multi scale feature fusion and class weight loss for skin lesion classification 2024
- [12] M A Javaid et al DSSCC Net enhanced skin cancer classification using SMOTE Tomek and optimized convolutional neural network 2025
- [13] S Bakheet et al A multi feature fusion framework for automatic skin cancer diagnostics 2023
- [14] P Murali and D H Mazumder DermaScanAI an explainable hybrid deep learning framework for automated skin lesion classification using dual attention and metadata fusion 2026
- [15] B Van Calster et al Calibration the Achilles heel of predictive analytics *BMC Medicine* vol 17 article 230 2019
- [16] P Tschandl et al Human computer collaboration for skin cancer recognition *Nature Medicine* vol 26 pp 1229 to 1234 2020
- [17] Z Li et al Artificial intelligence in dermatology image analysis current developments and future trends *Journal of Clinical Medicine* 2022
- [18] N Melarkode et al AI powered diagnosis of skin cancer a contemporary review open challenges and future research directions *Cancers* 2023