

Lightweight Explainable and Reliable Deep Learning for Multiclass Skin Lesion Classification

A Systematic Review of Metadata Fusion Class Imbalance Calibration and Leakage Safe Evaluation

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: Deep learning has improved automated skin lesion classification while practical clinical research still requires efficient models reliable confidence estimates interpretable predictions and evaluation designs that prevent information leakage; therefore this review examines lightweight architectures patient metadata fusion class imbalance strategies explainable artificial intelligence probability calibration and lesion grouped validation within multiclass dermoscopic classification; the review protocol covers major scientific databases with structured screening and evidence extraction while the current evidence map contains 68 deduplicated seed records with 30 priority studies and 29 completed abstract level extractions; preliminary synthesis shows rapid progress in convolutional and transformer models although direct comparison remains limited by inconsistent datasets split units class definitions metrics and reporting practices; moreover calibration and quantitative explanation assessment appear less common than accuracy reporting while lesion or patient grouped splitting is not always stated clearly; consequently future studies should combine class sensitive metrics transparent split governance calibration analysis explanation validation and measured computational efficiency within one reproducible protocol.

Index Terms: skin lesion classification; dermoscopy; lightweight deep learning; metadata fusion; class imbalance; explainable artificial intelligence; calibration; data leakage; systematic review.

I INTRODUCTION

Skin cancer and other cutaneous disorders create a substantial diagnostic burden while dermoscopy offers magnified visual information that supports clinical assessment; moreover deep neural networks have demonstrated strong discrimination in benchmark studies and large scale comparisons although a high classification score does not alone establish safety reliability or general clinical value [1] [2].

Multiclass dermoscopic datasets present connected methodological difficulties because common benign categories dominate training sets while rare malignant or clinically important categories provide fewer examples; furthermore multiple images can belong to one lesion and an image level split can place related observations in both training and evaluation partitions which can produce optimistic estimates [3].

Recent studies increasingly use MobileNet EfficientNet attention mechanisms transformers feature fusion and knowledge distillation to improve predictive performance or reduce computational cost; however comparisons across publications remain uncertain when dataset versions class mappings augmentation policies split units and reported metrics differ [4] [5].

Patient metadata can add age sex and anatomical site information to an image model while imbalance aware objectives can reduce domination by frequent classes; additionally visual explanation methods can reveal influential image regions and calibration methods can measure whether predicted confidence corresponds with observed correctness [6] [7].

Existing reviews describe artificial intelligence for dermatology or skin cancer detection while fewer syntheses examine efficiency metadata imbalance explainability calibration and leakage prevention as one connected evidence problem; therefore this review develops an integrated taxonomy and a reporting framework for reproducible multiclass skin lesion classification [8] [9] [10].

A Review Questions

This review asks which lightweight model families are used for multiclass dermoscopic classification; how patient metadata and class imbalance are handled; how explanation quality probability calibration computational efficiency and leakage safe evaluation are measured; and which combinations remain insufficiently studied.

II REVIEW METHODOLOGY

A Protocol and Reporting Approach

The review follows a predefined protocol that separates search screening extraction quality appraisal and synthesis; consequently the final manuscript will report the search flow with a PRISMA style diagram while every exclusion at full text stage will receive a recorded reason [11].

B Information Sources and Search Strategy

The planned information sources are Scopus Web of Science PubMed IEEE Xplore ScienceDirect SpringerLink and Google Scholar for citation chaining; the search combines skin lesion melanoma dermoscopy classification deep learning HAM10000 ISIC explainability lightweight metadata calibration uncertainty and leakage related concepts while database specific syntax will be preserved in the supplement.

C Eligibility Criteria

Eligible primary studies examine automated dermoscopic skin lesion classification with a sufficiently described dataset model and evaluation procedure; additionally each study must report outcomes that support critical assessment while editorials abstracts segmentation only articles unverifiable datasets duplicate versions and unrelated tasks are excluded.

D Study Selection and Extraction

Records are deduplicated by digital object identifier title and author year before title and abstract screening; subsequently eligible full texts are assessed while extracted fields cover dataset size class distribution split unit model architecture metadata fusion imbalance treatment metrics explainability calibration efficiency external validation statistical evidence and reproducibility resources.

E Quality Appraisal and Synthesis

Quality appraisal examines lesion or patient grouping test isolation class sensitive reporting calibration explanation evaluation efficiency measurement statistical uncertainty and availability of code splits or predictions; synthesis is primarily narrative because heterogeneous datasets and evaluation protocols can make a pooled performance estimate misleading.

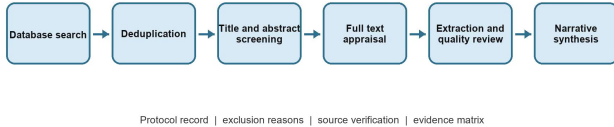


Fig 1 Structured review workflow used for evidence collection and synthesis

III PRELIMINARY EVIDENCE PROFILE

The working library contains 68 deduplicated seed papers while 30 directly relevant records have been prioritized and 29 records have completed abstract level extraction; however full text appraisal remains in progress so these values describe the current evidence base and will be replaced by the final screening counts before submission.

A Datasets and Evaluation Practice

HAM10000 remains a frequent benchmark because it provides 10015 dermoscopic images across seven diagnostic categories with lesion identifiers and limited patient metadata; nevertheless reported results can differ substantially when studies use random image splits instead of lesion grouped partitions or when preprocessing and class mappings are not comparable [3].

ISIC challenge data and other public collections broaden the evidence base while external validation remains important because acquisition devices clinical settings and population characteristics can change image appearance; therefore a strong internal benchmark should not be interpreted as proof of transportability to routine care [12].

B Model Families and Efficiency

Convolutional transfer learning remains common because pretrained backbones provide strong representations with manageable training demands; meanwhile MobileNet uses efficient operations for constrained hardware and EfficientNet balances network depth width and resolution which makes both families relevant to deployment oriented comparisons [13] [14].

Attention mechanisms hybrid convolutional transformer models pruning and knowledge distillation extend the efficiency design space; however parameter count alone does not establish practical efficiency because latency memory preprocessing cost runtime software and energy consumption also influence deployment behavior [15] [16].

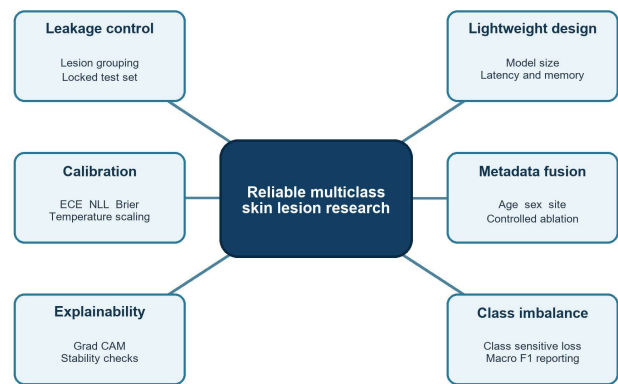


Fig 2 Integrated taxonomy for reliable multiclass skin lesion research

C Metadata Fusion

Age sex and anatomical location can provide context that is not present in dermoscopic pixels while fusion may occur before classification through concatenation gating or attention; nevertheless metadata preprocessing must be fitted on training data only and missing or unknown values require explicit handling to prevent optimistic evaluation.

Metadata benefit should be demonstrated through controlled image only and image plus metadata ablations using an unchanged split and training protocol; otherwise improvements cannot be attributed confidently to the added clinical variables.

D Class Imbalance

Skin lesion collections are strongly imbalanced because melanocytic nevi often outnumber rarer categories; consequently accuracy can conceal weak minority class recognition while macro F1 balanced accuracy per class sensitivity and confusion matrices provide more informative evidence [17].

Common responses include augmentation oversampling synthetic sampling class weighting focal objectives and class balanced losses; however aggressive resampling can distort prevalence while loss modifications require controlled ablation because an overall gain may coexist with poorer calibration or unstable minority class behavior.

E Explainability

Grad CAM and related activation mapping methods are widely used because they provide visual localization without changing the trained architecture; however a visually plausible heatmap is not a clinical validation result and explanation evidence becomes stronger when stability localization sensitivity or expert agreement is measured [6].

F Calibration and Uncertainty

Calibration asks whether predicted confidence agrees with empirical correctness while discrimination asks whether classes are separated; therefore expected calibration error negative log likelihood and Brier score should accompany accuracy measures when model confidence may influence a decision support workflow [7] [18].

Temperature scaling offers a simple post training correction when its parameter is fitted on validation predictions and the locked test set remains untouched; nevertheless calibration can change across populations which means external assessment remains necessary.

G Leakage and Reproducibility

Repeated views of one lesion create a grouped data structure that must be respected during partitioning; moreover preprocessing feature selection hyperparameter tuning calibration threshold choice and model selection must remain isolated from the final test set.

Reproducibility improves when authors publish exact dataset identifiers lesion grouped split manifests random seeds software versions model checkpoints and prediction files; additionally confidence intervals and paired tests should reflect clustering when several images belong to the same lesion.

H Comparative Performance Interpretation

Performance values reported by different studies should not be ranked without checking whether the same dataset release diagnostic labels split unit and evaluation set were used; furthermore a model tested on a random image partition can obtain an apparently stronger result than a model tested with lesion grouping even when its real generalization is weaker.

Accuracy remains useful as a descriptive measure while macro F1 and balanced accuracy provide a clearer summary when class frequencies are unequal; moreover per class sensitivity and precision are necessary because a high overall score can coexist with poor recognition of melanoma dermatofibroma vascular lesions or other uncommon categories.

Area under the receiver operating characteristic curve is frequently reported although it can appear favorable in imbalanced settings and does not identify an operating threshold; therefore decision threshold selection precision recall behavior and calibration should be interpreted together rather than treating one scalar metric as complete evidence.

I External Validation and Clinical Relevance

External validation tests whether a model retains performance when the acquisition source patient population device prevalence or clinical pathway changes; consequently multi dataset experiments provide stronger transportability evidence than a single internal split although harmonization choices and label mapping must remain transparent [19] [20].

Clinical relevance also depends on the intended use because triage decision support educational assistance and autonomous diagnosis require different evidence thresholds; therefore review conclusions must distinguish technical classification performance from prospective clinical utility and must not describe benchmark success as regulatory or clinical validation.

J Fairness and Population Representation

Skin tone age sex and anatomical site can influence image appearance disease prevalence and model behavior while many public dermoscopic datasets provide limited demographic information; consequently subgroup performance and dataset representation should be reported whenever the available metadata permits a responsible analysis [21].

Metadata fusion can improve contextual reasoning while it can also encode site specific prevalence or demographic shortcuts; therefore subgroup assessment missingness analysis and external validation are required before metadata benefit is interpreted as universally reliable.

IV INTEGRATED RESEARCH GAPS

The preliminary synthesis indicates that many studies optimize classification performance while fewer evaluate all six dimensions within one controlled protocol; specifically lightweight design is often reported without measured resource cost while explainability is frequently illustrated without quantitative assessment and calibration is less visible than discrimination.

A defensible research agenda should therefore combine lesion grouped evaluation metadata ablation imbalance aware learning class sensitive metrics post training calibration quantitative explanation checks and direct efficiency measurements; furthermore external validation and subgroup analysis are required before claims extend beyond the development dataset.

Abstract level evidence map for eight priority studies

	Metadata	Imbalance	XAI	Calibration	Efficiency	External data	Grouped split
DermaScanAI	Reported	NR	Reported	Reported	Reported	NR	NR
Arshad et al	NR	Reported	Reported	NR	Reported	NR	NR
Abdulredah et al	NR	NR	Reported	NR	NR	Reported	NR
DBSCC Net	NR	Reported	Reported	NR	NR	Reported	NR
Sayed et al	Reported	NR	Reported	NR	Reported	Reported	NR
Pavel et al	NR	NR	Reported	NR	Reported	NR	NR
Yadav et al	NR	NR	NR	NR	Reported	NR	NR
Hu et al	NR	Reported	NR	NR	NR	Reported	NR

Fig 3 Preliminary abstract level evidence gap map where NR means not reported in the abstract

A Gap in Joint Evaluation

The central gap is not the absence of high performing classifiers because many architectures already report strong benchmark results; instead the gap concerns the limited joint evaluation of efficiency metadata contribution minority class behavior calibrated confidence explanation reliability and leakage control under one fixed protocol.

B Gap in Reproducible Comparison

Reproducible comparison remains difficult when publications omit lesion identifiers exact partitions random seeds checkpoints or prediction files; moreover incomplete reporting prevents independent calculation of confidence intervals error overlap and paired significance tests.

C Gap in Explanation Validation

Explanation research often presents a small number of favorable heatmaps while negative examples unstable maps and clinically irrelevant attention receive less discussion; consequently future work should define explanation sampling rules and quantitative checks before reviewing the visual outputs.

V RECOMMENDED REPORTING FRAMEWORK

Future studies should report dataset provenance inclusion rules duplicate handling patient or lesion grouping class distributions and locked test governance; they should also document preprocessing augmentation metadata encoding model provenance optimization settings and the exact decision rule used for selection.

Results should include macro F1 balanced accuracy per class measures confusion matrices calibration metrics confidence intervals and resource measurements; additionally explanation methods should be evaluated beyond selected examples while code split manifests and predictions should be preserved wherever licensing permits.

A Data and Split Reporting

Authors should identify the dataset version download date license image count lesion count patient count class mapping duplicate policy and missing metadata treatment; furthermore the split table should report images and unique groups per partition so leakage checks can be reproduced.

B Training and Selection Reporting

The training record should include initialization optimizer learning rate schedule batch size epochs stopping rule augmentation loss function sampling policy and random seed; additionally the manuscript should name the primary validation metric and preserve the final test set until every design choice is frozen.

C Reliability and Efficiency Reporting

Reliability reporting should include raw and calibrated confidence measures with the calibration fitting partition identified explicitly; meanwhile efficiency reporting should include parameter count model size latency memory and hardware or runtime details so claims remain interpretable.

D Explanation Reporting

Explanation reporting should identify the target layer algorithm parameters case selection procedure and any normalization applied to heatmaps; moreover authors should include failures and stability evidence so readers can distinguish representative behavior from illustrative examples.

VI DISCUSSION

The reviewed evidence suggests that progress in dermoscopic classification is broad yet fragmented because architectural innovation often advances faster than evaluation discipline; therefore future value is likely to come from integrated protocols that measure discrimination reliability interpretability and efficiency with equal transparency.

Lightweight architectures can support deployment on limited hardware while reduced size should not be treated as an automatic advantage; rather a compact model should demonstrate acceptable minority class performance stable confidence and measured runtime under the intended operating environment.

Metadata fusion is promising because clinical context influences diagnostic reasoning while its contribution must be tested through ablation and subgroup analysis; similarly imbalance methods should be evaluated for their effect on both class discrimination and calibration instead of being selected from accuracy alone.

Explainable artificial intelligence can support error analysis and communication while current heatmaps do not establish causal reasoning or clinical safety; consequently explanation results should remain one component of a broader validation framework that includes robustness calibration and human evaluation.

The proposed reporting framework can improve comparison across future studies because it connects dataset governance model development statistical evaluation and deployment evidence; however adoption also requires journals reviewers and dataset curators to request group aware partitions and transparent artifacts.

VII LIMITATIONS

The present manuscript is an active review draft because final database searches and full text appraisal are not yet complete; consequently the preliminary synthesis can guide the taxonomy and research questions while definitive study counts quality findings and comparative conclusions must wait for the frozen evidence matrix.

Publication bias incomplete reporting inaccessible full texts and inconsistent terminology may limit the final synthesis; moreover rapid model development can make a fixed search outdated which requires a documented final search update immediately before submission.

VIII CONCLUSION

Deep learning for multiclass skin lesion classification is progressing from conventional convolutional models toward efficient multimodal and explainable systems; however trustworthy evidence depends on leakage safe evaluation class sensitive reporting calibrated confidence validated explanations and transparent resource measurement while the final review will use completed full text appraisal to determine how consistently these requirements are met.

TABLE I INTEGRATED REVIEW DIMENSIONS AND CRITICAL QUESTIONS

Dimension	Evidence to extract	Critical question
Efficiency	Parameters size latency memory energy	Is efficiency measured on a stated platform
Metadata	Variables encoding missingness fusion ablation	Does metadata provide controlled incremental value
Imbalance	Sampling losses augmentation class metrics	Are minority classes evaluated explicitly
Explainability	Method samples stability localization review	Are explanations evaluated beyond selected images
Calibration	ECE NLL Brier scaling uncertainty	Does confidence correspond with correctness
Leakage control	Group unit test isolation split availability	Can related observations cross partitions

TABLE II REPRESENTATIVE PRIORITY STUDIES FROM THE CURRENT EVIDENCE MATRIX

Study	Data	Main focus	Evidence visible in abstract	Key appraisal need
Murali et al 2026	HAM10000	Attention and metadata fusion	XAI calibration efficiency	Verify grouped split and full protocol
Arshad et al 2025	HAM10000 and ISIC 2018	Network fusion and XAI	Imbalance XAI efficiency	Verify split unit and calibration
Abdulredah et al 2025	HAM10000 and ISIC data	Deep feature fusion	XAI and external data	Verify imbalance and resource measures
Javaid et al 2025	HAM10000 ISIC 2018 and PH2	SMOTE Tomek with CNN	Imbalance XAI external data	Verify grouped split and calibration

Study	Data	Main focus	Evidence visible in abstract	Key appraisal need
Sayeed et al 2026	HAM10000 and ISIC 2019	Federated lightweight CNN	Metadata XAI efficiency external data	Verify uncertainty and split unit
Pavel et al 2025	HAM10000	Knowledge distillation	XAI and efficiency	Verify calibration and external validity
Yadav et al 2024	HAM10000	Lightweight cross attention	Efficiency	Verify class imbalance and leakage control
Hu et al 2024	HAM10000 and ISIC 2019	Feature fusion and class weighting	Imbalance and external data	Verify XAI calibration and grouping

TABLE III MINIMUM REPORTING CHECKLIST FOR FUTURE STUDIES

Domain	Minimum information	Reason
Dataset	Version source license images lesions patients classes missingness	Establish provenance and population structure
Split	Group unit seed manifests counts and leakage audit	Prevent related observations crossing partitions
Training	Initialization optimizer schedule epochs stopping augmentation and loss	Support reproduction and fair ablation
Selection	Primary validation metric and locked test governance	Prevent test guided tuning
Performance	Macro F1 balanced accuracy per class results and confusion matrix	Show minority class behavior
Statistics	Confidence intervals paired tests and clustering method	Quantify uncertainty and comparison strength
Calibration	ECE NLL Brier method and fitting partition	Assess confidence reliability
XAI	Method target layer sampling failures and stability checks	Move beyond selected visual examples
Efficiency	Parameters size latency memory hardware and runtime	Support deployment claims
Artifacts	Code splits checkpoints predictions and environment	Enable independent verification

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 58 dermatologists. *Annals of Oncology* vol 29 no 8 pp 1836 to 1842. 2018
- [3] P Tschandl C Rosendahl and H Kittler. The HAM10000 dataset: a large collection of multisource dermatoscopic images of common pigmented skin lesions. *Scientific Data* vol 5 article 180161. 2018
- [4] Z Hu W Mei H Chen and W Hou. Multi scale feature fusion and class weight loss for skin lesion classification. 2024
- [5] D P Yadav et al. Dual scale lightweight cross attention transformer for skin lesion classification. 2024
- [6] R R Selvaraju et al. Grad CAM visual explanations from deep networks via gradient based localization. in *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. in *Proceedings of the International Conference on Machine Learning*. 2017
- [8] Z Li et al. Artificial intelligence in dermatology image analysis current developments and future trends. *Journal of Clinical Medicine*. 2022

- [9] N Melarkode K Srinivasan S M Qaisar and P Plawiak AI powered diagnosis of skin cancer a contemporary review open challenges and future research directions *Cancers* 2023
- [10] B C R S Furriel et al Artificial intelligence for skin cancer detection and classification for clinical environment a systematic review 2024
- [11] M J Page et al The PRISMA 2020 statement an updated guideline for reporting systematic reviews *BMJ* vol 372 2021
- [12] P Tschandl et al Human computer collaboration for skin cancer recognition *Nature Medicine* vol 26 pp 1229 to 1234 2020
- [13] A Howard et al Searching for MobileNetV3 in *Proceedings of the IEEE International Conference on Computer Vision* 2019
- [14] M Tan and Q V Le EfficientNet rethinking model scaling for convolutional neural networks in *Proceedings of the International Conference on Machine Learning* 2019
- [15] Y Nie P Sommella M Carratu M O Nils and J Lundgren A deep CNN transformer hybrid model for skin lesion classification of dermoscopic images using focal loss 2022
- [16] M A Pavel et al Multi stage knowledge distillation with layer fusion based deep learning approach for skin cancer classification 2025
- [17] M A Javaid et al DSSCC Net enhanced skin cancer classification using SMOTE Tomek and optimized convolutional neural network 2025
- [18] B Van Calster et al Calibration the Achilles heel of predictive analytics *BMC Medicine* vol 17 article 230 2019
- [19] C Hernandez Perez et al BCN20000 dermoscopic lesions in the wild *Scientific Data* 2024 doi 10.1038/s41597-024-03387-w
- [20] Y Liu et al A deep learning system for differential diagnosis of skin diseases *Nature Medicine* vol 26 pp 900 to 908 2020
- [21] R H Patel E A Foltz A Witkowski and J Ludzik Analysis of artificial intelligence based approaches applied to non invasive imaging for early detection of melanoma a systematic review 2023
- [22] M Mateen S Hayat F Arshad Y H Gu and M A Al Antari Hybrid deep learning framework for melanoma diagnosis using dermoscopic medical images 2024
- [23] S Bibi et al MSRNet multiclass skin lesion recognition using additional residual block based fine tuned deep models information fusion and best feature selection 2023
- [24] A Naeem and T Anees DVFNet a deep feature fusion based model for the multiclassification of skin cancer utilizing dermoscopy images 2024
- [25] H Cheng J Lian and W Jiao Enhanced MobileNet for skin cancer image classification with fused spatial channel attention mechanism 2024
- [26] S Bakheet S Alsubai A El Nagar and A Alqahtani A multi feature fusion framework for automatic skin cancer diagnostics 2023
- [27] A H Efat S M Hasan M P Uddin and M Mamun A multi level ensemble approach for skin lesion classification using customized transfer learning with triple attention 2024
- [28] M Hussain et al SkinNet INIO multiclass skin lesion localization and classification using fusion assisted deep neural networks and improved nature inspired optimization algorithm 2023
- [29] M Azeem K Kiani T Mansouri and N Topping SkinLesNet classification of skin lesions and detection of melanoma cancer using a novel multilayer deep convolutional neural network 2023
- [30] A Ray F Sarkar F Schwenker and R Sarkar Decoding skin cancer classification perspectives insights and advances through researchers lens 2024