An Uncertainty-Aware Hybrid Multimodal Deep Learning Framework for Skin Cancer Detection

Sonia Arora[1], Shashi Kant Gupta[2,3]
[1,2] Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan, Malaysia;
[3]Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, 140401, Punjab, Índia.
sonia.phd21@gmail.com , shashigupta@lincoln.edu.my
ORCID: [1] https://orcid.org/0009-0004-1833-3951, [2] https://orcid.org/0000-0001-6587-5607

Abstract: While automatic classification of skin lesions for computer aided assessment in dermatology has been proven to be successful, it is not sufficient to assume that it is good image classification performance, to achieve good clinical decision support. A hybrid multimodal deep learning approach that fuses dermoscopic appearance, clinical metadata and predictive uncertainty is proposed in this paper. The data set contains 10015 dermoscopic images of 7 diagnostic categories, age, gender and location of the lesion (HAM10000). Images are normalised, enhanced and metadata is cleaned, encoded and normalised. The feature extraction module used for the visual feature is based on efficientNetV2B0, the clinical information is expressed by a clinical encoder, and the attention/gated mechanism is used to integrate the two modalities. To estimate the predictive uncertainty, Monte Carlo Dropout is kept on for inference. The splitting of patients/lesions, controlled ablations, macro-averaged measures, calibration and uncertainty analysis are the foundations of the validation design. Two goals of the study are to see if clinical information helps to improve class-sensitive performance and to see if uncertainty estimates can detect potentially unreliable predictions.
Keywords: HAM10000; Dermoscopy; Clinical Metadata; Monte Carlo Dropout; Patient-Level Validation.

1. Introduction

Skin cancer is an important public-health issue because the benign and malignant growths (such as melanoma) can spread rapidly when not detected early. Dermoscopy gives the dermatologist a wealth of information about the internal components of the lesion including colour, texture, shape and the border. Interpretation can be difficult, however, since various categories of lesions can have similar visual patterns. It has stimulated the progress of computer-aided diagnosis systems for machine learning and deep learning [1, 2]. The deep learning models can be trained with hierarchical visual representation from dermoscopic images and transfer learning is applicable for the medical imaging tasks, too [3,4]. When adding more clinical branches and uncertainty branches, efficient architectures are effective, e.g., EfficientNetV2 is an efficient architecture with good representation capability [5]. However, there are a few aspects that are not reflected in dermoscopic images that are significant for clinical assessment. HAM10000 consists of 10,015 images from 7 diagnostic categories, and has age, sex and lesion-localization metadata [1] available for further investigation of complementary visual and clinical information. More recently, there have been multimodal studies indicating that clinical information can be utilized alongside image features, but a simple stringing together of all the image features may be insufficient to account for the differing importance of each modality in the different types of lesions. Another way is to learn the contribution of each modality either by paying attention to it or by gating it in. A further challenge is the reliability of the prediction: a prediction can be very confident for an ambiguous input. The uncertainty in the prediction can be quantified by numerous times running stochastic inference with Monte Carlo Dropout [4]. Evaluation should also prevent the leakage of information. There could be several images in the lesion and false generalisation estimates can be caused by inappropriate image level splitting. In this way, the value of patient/lesion-level separation in the presence of identifiers becomes the focus of this study, along with the focus shifting from accuracy to the following: macro-F1, class-wise recall, ROC-AUC, calibration, confusion matrices, and uncertainty. Proposed framework is seven class classification based on the fusion of the visual features of the EfficientNetV2B0 [5] with the clinical encoder, attention/gated fusion and MC Dropout [4] along with controlled experiments of image only, conventional concatenation and gated fusion.

2. Related Work

Research on automated skin-lesion classification has progressed from image-only deep learning toward multimodal and reliability-aware systems. Image-based studies have investigated CNN–Transformer hybrids, machine-learning diagnostic tools, deep convolutional networks, and augmentation/transfer-learning classifiers [6–9]. These approaches demonstrate the value of learned visual representations but generally do not exploit clinical metadata as an explicit complementary modality.

2.1 Multimodal Learning Using Clinical Metadata

Adebiyi et al. investigated multimodal learning on HAM10000 [10]. Das et al. compared concatenation, weighted fusion, self-attention, and cross-attention architectures, showing that fusion strategy is an important design choice [11]. Tran-Van and Le investigated cross-attention to model relationships between image and metadata representations [12], while Aboulmira et al. explored attention-guided multimodal learning with reliability-oriented fusion [13]. These studies indicate that clinical metadata can complement dermoscopic information, while the fusion mechanism can influence performance and computational complexity.

2.2 Reliability and Advanced Classification

Uncertainty-aware prediction is important because confidence alone does not guarantee correctness. MC Dropout provides a practical uncertainty-estimation mechanism [4], while recent studies have investigated additional methods for robust representation learning, including stepwise self-knowledge distillation [14] and attention-based optimized classification [15]. However, multimodal learning, explicit uncertainty estimation, and leakage-aware validation are not consistently evaluated together.

2.3 Research Need

The resulting gap is a need for a controlled framework that jointly considers visual information, clinical context, predictive uncertainty, and credible validation. The present study addresses this gap through

EfficientNetV2B0 visual extraction, a clinical encoder, attention/gated fusion, MC-Dropout uncertainty, and patient/lesion-level validation. The goal is to assess not only classification performance but also whether multimodal fusion and uncertainty provide measurable reliability benefits.

Table 1. Comparison of this work with representative studies on skin-lesion classification

Study	Input	Architecture / approach	Fusion	Uncertainty	Major focus
Tschandl et al. [1]	Images + metadata	HAM10000 benchmark	Dataset-level	No	Large multiclass benchmark with clinical information
Yap et al. [2]	Images + clinical data	Deep multimodal learning	Image–metadata	No	Demonstrated multimodal classification potential
Adebiyi et al. [10]	Images + metadata	Multimodal learning	Multimodal	No	HAM10000 multimodal classification
Das et al. [11]	Images + metadata	Comparative multimodal models	Concat./weighted/attention	No	Compared fusion strategies
Tran-Van & Le [12]	Images + metadata	Cross-attention model	Cross-attention	No	Learned cross-modal interactions
Aboulmira et al. [13]	Images + metadata	EfficientNet multimodal	Attention/evidential	Reliability-oriented	Adaptive multimodal reliability
This work	Images + metadata	EfficientNetV2B0 + clinical encoder	Attention/gated	MC Dropout	Leakage-aware validation, ablation, macro-F1 and uncertainty

3. Method, Experiments and Results

The proposed framework categorizes the seven types of lesions through two-stage learning: First through dermoscopic images and secondly through clinical metadata. The workflow includes data preparation, feature extraction based on the modality, multimodal fusion, classification with uncertainty assessment, and leakage-aware validation. The figure 1 shows the proposed methodology consists of eight stages: (1) Image Preprocessing, where dermoscopic images are resized, normalized, and augmented to improve robustness; (2) Metadata Preprocessing, where clinical data are cleaned, encoded, and normalized; (3) Feature Extraction, where EfficientNetV2B0 extracts image features and a dedicated encoder extracts clinical features; (4) Hybrid Feature Fusion, where both feature representations are integrated using an attention/gated mechanism; (5) Fully Connected Classification, where the fused features are used for seven-class skin-lesion classification; (6) Monte Carlo Dropout for Uncertainty, where multiple stochastic predictions estimate prediction uncertainty; (7) Skin Cancer Classification and Confidence Score, where the model provides the predicted lesion class along with its confidence; and (8) Performance Evaluation and Patient-Level Validation, where the model is evaluated using Accuracy, Precision, Recall, F1-score, ROC-AUC, and Confusion Matrix under patient-level data splitting.

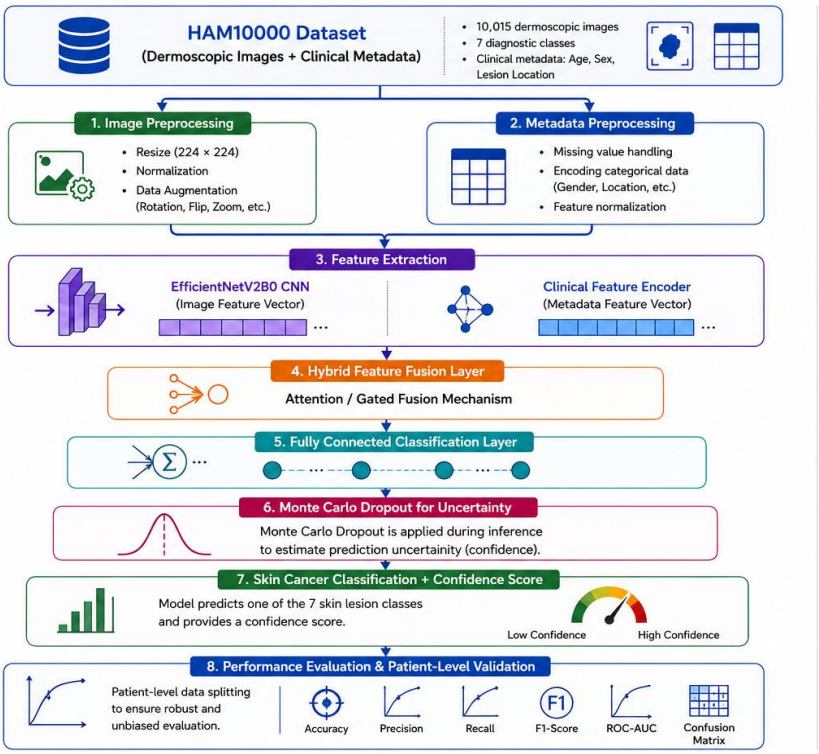

Figure 1. Proposed Uncertainty-Aware Multimodal Framework for Skin-Lesion Classification

3.1 Method

HAM10000 is a collection of 10,015 RGB dermoscopic images of actinic keratosis/intraepithelial (akiec), basal cell (bcc), benign keratosis-like (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (vasc), as well as age, sex and lesion-localization metadata [1]. Images are only resized, normalized and augmented during training. Cleaning, encoding and normalization of clinical variables is performed based on information in the training sets only.

3.1.1 Visual Feature Extraction

The visual backbone is chosen to be EfficientNetV2B0, which offers efficient hierarchical representation learning [5]. The image feature vector is defined by equation 1:

$$F_I = f(X; W) \quad (1)$$

where X is the input dermoscopic image, W denotes visual-network parameters, and F_I is the extracted image representation.

3.1.2 Clinical Feature Representation

A separate clinical encoder transforms structured metadata into a numerical representation by equation 2:

$$F_C = g(M; \theta) \quad (2)$$

where M represents clinical metadata and θ denotes clinical-encoder parameters. Keeping the branches separate until fusion allows each modality to learn a representation suited to its structure.

3.1.3 Attention/Gated Multimodal Fusion

The visual and clinical representations are integrated through a learned attention/gated mechanism rather than assuming equal contribution. A simplified formulation as shown by equation 3:

$$F_F = \alpha F_I + (1 - \alpha) F_C \quad (3)$$

where F_F is the fused representation and α is a learned modality-weighting factor. The fused representation is passed to fully connected layers and a seven-class Softmax classifier. The design is motivated by multimodal studies showing that fusion strategy influences performance [10–13].

3.1.4 Uncertainty Estimation

MC Dropout is retained during inference to generate T stochastic predictions for each input [4]. The predictive mean is given by equation 4:

$$\hat{y}(x) = (1/T)\ \Sigma^{t}_{=1}{}^{T}\ f_t(x) \qquad (4)$$

where $f_t(x)$ denotes the prediction from the t-th stochastic forward pass. Variation among stochastic predictions is used as an uncertainty signal; greater variation indicates less stable prediction.

3.2 Experimental Design

The evaluation is done for three configurations: (A) an image-only EfficientNetV2B0 baseline, (B) a multimodal baseline using conventional concatenated image and clinical representations, and (C) the proposed model with attention/gated fusion and MC-Dropout uncertainty. This design is based on the idea of isolating the contribution of clinical metadata and then comparing the learned fusion to simple concatenation.

3.3 Data Splitting and Leakage Control

Wherever possible, patient and/or lesion identifiers will be provided to allow correlated samples to stay in the same partition as multiple images may describe the same lesion. Normalization, how to deal with missing values, and how to map a categorical value to a single number, are determined from the training data. This procedure aims to prevent the leakage of information and to give a better approximation of generalization.

3.4 Evaluation Metrics

The metrics used for evaluating their performance include accuracy, precision, recall, macro-F1, ROC-AUC, confusion matrices, calibration/reliability measures, and MC-Dropout uncertainty. The emphasis is given to overall accuracy and macro-F1 and class-specific recall since low accuracy on less frequent diagnostic categories may be present.

3.5 Results and Expected Experimental Findings

The planned analysis uses the same leakage-aware split for all three configurations. Complementary or better macro-F1 or minority-class recall in multimodal models would support the inclusion of complementary clinical information; better gated fusion than concatenation would support adaptive modality weighting. The uncertainty analysis will analyze if the error in the predictions has more stochastic variation. If split at the image level were also evaluated, the effect of leakage control would be quantified using any reduction in splitting done at the patient/lesion level.

4. Observation and Discussion

The proposed study will aim to go beyond the traditional goal of a high overall classification accuracy. The contribution of clinical metadata, effectiveness of multimodal fusion, behavior of uncertainty estimates, and reliability of the validation strategy are the main issues that are focused on. The multimodal experiments are suggested and not actually carried out, so the following discussion presents observations to be considered rather than the unsupported numerical results.

4.1 Observation of Image-Based and Multimodal Learning

The image-only EfficientNetV2B0 baseline will be compared to the multimodal models in the first observation. The image-only model gives a reference to evaluate the amount of diagnostic information that can be obtained from the appearance of the image. Age, sex, and lesion localization from HAM10000 are added to the multimodal approach. When the multimodal models outperform in terms of macro-F1 or class-specific recall, especially in cases where the categories share similar images, it means that clinical information adds further evidence beyond what is encountered in the image. In addition to the overall accuracy, the accuracy of individual classes should be evaluated based on recall, precision and macro F1.

4.2 Observation of Fusion Strategy

The second observation relates to the use of conventional feature concatenation instead of attention/gated fusion. Simple concatenation is used to combine image and clinical representations, without making an explicit attempt to learn the relative importance of each modality; gated fusion is designed to learn the relative importance of each representation. Gated fusion outperforming

concatenation would indicate that there is no simple feature joining that is sufficient to capture all of the complementary aspects of a relationship. On the other hand, if the performance gap was small, it would suggest that the added complexity of learned fusion might not be very useful in practice. This comparison helps to avoid the justifying of complexity in architecture without evidence.

4.3 Observation of Minority-Class Performance

In HAM10000, class imbalance is an important factor. It is important to look at the confusion matrix and class-wise recall carefully because a model may be able to obtain good overall accuracy and perform poorly on the less common categories. Special consideration should be made for categories like melanoma and other clinically important categories that are not as well represented. Better macro-F1 and minority-class recall would be good proof of improvement than accuracy would be. Also, it should identify if the errors are still within certain visual category. The addition of metadata may not be enough to resolve some of the classes, in which case it could be due to the lack of discriminative information in the available clinical variables.

4.4 Observation of Predictive Uncertainty

MC Dropout generates multiple stochastic predictions for the same input, with their variation used as an estimate of predictive uncertainty. The key observation will be whether incorrect predictions tend to exhibit higher uncertainty than correct predictions. If such a relationship is observed, uncertainty could provide a warning signal for ambiguous cases. For example, similar probabilities for two diagnostic categories may indicate greater uncertainty than repeated predictions that consistently favour one category. Nevertheless, high uncertainty should not automatically be interpreted as evidence of an incorrect prediction; it must be evaluated quantitatively and calibrated against actual outcomes. Uncertainty is therefore treated as an additional measurable outcome rather than a replacement for standard classification metrics.

4.5 Observation of Data Leakage and Generalization

HAM10000 contains multiple images associated with the same lesion, creating a potential leakage pathway when related samples are randomly distributed across training and testing sets. Patient/lesion-level splitting is therefore expected to provide a more realistic estimate of generalization. If performance decreases compared with conventional image-level random splitting, this should not necessarily be interpreted as a weakness of the proposed model; it may indicate that image-level evaluation benefited from similarities between related samples. A conservative estimate obtained through leakage-aware validation is scientifically preferable because inflated scores can give a misleading impression of real-world performance.

5. Conclusion

This study proposes a method to classify skin lesions into seven classes with uncertainty assessment based on the dermoscopic images and clinical metadata. It can provide visual representations for efficientNetV2B0, a dedicated clinical encoder for structured information, and attention/gated fusion combining the modalities. Patient/lesion-level splitting and class-sensitive metrics enhance evaluation; MC Dropout provides predictive uncertainty. The controlled design separates out the contribution of clinical metadata from the fusion strategy and does not presume that multimodal learning necessarily provides benefits. The framework will be tested on final experiments to see if it enhances the class sensitive classification and if the uncertainty estimation can detect potentially incorrect predictions.

References

[1]. Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data, 5, 180161. https://doi.org/10.1038/sdata.2018.161

[2]. Yap, J., Yolland, W., & Tschandl, P. (2018). Multimodal skin lesion classification using deep learning. Experimental Dermatology, 27(11), 1261–1267. https://doi.org/10.1111/exd.13777

[3]. Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., Paoli, J., Puig, S., Rosendahl, C., Soyer, H. P., Zalaudek, I., & Kittler, H. (2020). Human–

computer collaboration for skin cancer recognition. Nature Medicine, 26(8), 1229–1234. https://doi.org/10.1038/s41591-020-0942-0
[4]. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (Vol. 48, pp. 1050–1059). PMLR.
[5]. Tan, M., & Le, Q. V. (2021). EfficientNetV2: Smaller models and faster training. In Proceedings of the 38th International Conference on Machine Learning (Vol. 139, pp. 10096–10106). PMLR.
[6]. Nie, Y., Sommella, P., Carratù, M., O'Nils, M., & Lundgren, J. (2023). A deep CNN transformer hybrid model for skin lesion classification of dermoscopic images using focal loss. Diagnostics, 13(1), 72. https://doi.org/10.3390/diagnostics13010072
[7]. Tajerian, A., Kazemian, M., Tajerian, M., & Akhavan Malayeri, A. (2023). Design and validation of a new machine-learning-based diagnostic tool for the differentiation of dermatoscopic skin cancer images. PLOS ONE, 18(4), e0284437. https://doi.org/10.1371/journal.pone.0284437
[8]. Houssein, E. H., Abdelkareem, D. A., Hu, G., Abdel Hameed, M., Ibrahim, I. A., & Younan, M. (2024). An effective multiclass skin cancer classification approach based on deep convolutional neural network. Cluster Computing, 27, 12799–12819. https://doi.org/10.1007/s10586-024-04540-1
[9]. LesNet: An automated skin lesion deep convolutional neural network classifier through augmentation and transfer learning. (2024). Procedia Computer Science, 235, 112–121. https://doi.org/10.1016/j.procs.2024.04.014
[10]. Adebiyi, A., Abdalnabi, N., Smith, E. H., Hirner, J., Simoes, E. J., Becevic, M., & Rao, P. (2024). Accurate skin lesion classification using multimodal learning on the HAM10000 dataset. medRxiv. https://doi.org/10.1101/2024.05.30.24308213
[11]. Das, A., Agarwal, V., & Shetty, N. P. (2025). Comparative analysis of multimodal architectures for effective skin lesion detection using clinical and image data. Frontiers in Artificial Intelligence, 8. https://doi.org/10.3389/frai.2025.1608837
[12]. Tran-Van, N.-Y., & Le, K.-H. (2025). A multimodal skin lesion classification through cross-attention fusion and collaborative edge computing. Computerized Medical Imaging and Graphics, 124, 102588. https://doi.org/10.1016/j.compmedimag.2025.102588
[13]. Aboulmira, A., et al. (2025). Attention-guided multimodal skin disease classification using discounted Dempster–Shafer fusion and EfficientNet. Informatics in Medicine Unlocked, 58, 101699. https://doi.org/10.1016/j.imu.2025.101699
[14]. Zheng, J., Xie, K., Zhang, D., Lv, Z., & Yu, X. (2025). Stepwise self-knowledge distillation for skin lesion image classification. Scientific Reports, 15, 25238. https://doi.org/10.1038/s41598-025-10717-4
[15]. Priyadarsini, R. N., Tyagi, B., & Priyadharsini, M. (2026). An attention based optimized network for the classification of skin lesions. Scientific Reports, 16, 3494. https://doi.org/10.1038/s41598-025-31220-w