Original Article

Multimodal Skin Disease Classification via Gated Cross-Modal Fusion of Dermoscopic Images and Clinical Metadata

Chintan N. Kanani [1], Pradyumansinh U. Jadeja [1]

1. *Computer Science & Engineering, Darshan University, Rajkot, IND*

Abstract

The classification of skin diseases from dermoscopy images is an important step in the early diagnosis of skin diseases and in providing clinicians with decision support. While existing deep learning techniques use only visual data to make diagnostic decisions, they do not take into account other clinically relevant information such as a patient's gender, age, or location of the lesion. Because of the limitations of these unimodal systems, the robustness of a diagnostic decision can be severely compromised when attempting to diagnose skin diseases that are both visually ambiguous and underrepresented. DermaFusionNet is proposed in this paper as a novel multimodal deep learning architecture that combines dermoscopic image features with structured clinical metadata using a gate-based cross-modal fusion mechanism. The architecture described in this paper incorporates ConvNeXt visual feature extractors and dedicated metadata encoders, and it also employs an adaptive gated module to adjust the effect of each modality to ensure stable and contextually relevant feature integration across modalities. To further enhance the stability of multimodal training, the authors have included auxiliary modals to reduce the potential for one modality to dominate during training. The authors demonstrate the effectiveness of DermaFusionNet by evaluating it using the HAM10000 dataset and a stratified experimental protocol. The experimental results show that DermaFusionNet achieves an overall classification accuracy of 89.22%, with a precision of 86.32%, a recall of 89.22%, and an F1-score of 89.08%. Furthermore, DermaFusionNet outperformed not only the image-only baseline but also the naïve fusion baseline. The authors conclude that the use of gated multimodal fusion techniques has the potential to improve the reliability of automated skin disease classification systems.

Categories: AI applications, Machine Learning (ML), Deep Learning

Keywords: machine learning, deep learning, neural network, clinical metadata, ham10000 dataset

Introduction

Dermatological disorders can be considered some of the most significant health issues affecting people around the world. Skin cancers, especially melanoma, make up a large portion of the disease burden associated with dermatological disorders. In addition, the ability of early detection, proper treatment, and prompt clinical intervention to lead to improved patient outcomes through increased mortality and morbidity rates related to skin cancers is critical. Dermoscopy provides a method of visualizing and evaluating skin lesions through the use of infrared light, which allows for a much more detailed view of skin lesions than what can be accomplished using only visible light from the human eye. On the contrary, dermoscopy requires a high level of clinical experience, knowledge, and expertise, and because of this, variability in the diagnostic processes of the same dermoscopic picture can be problematic for everyone involved.

Automated classification systems for skin diseases using artificial intelligence (AI) and deep learning have been extensively explored to address the challenges associated with manual diagnosis. Convolutional neural networks (CNNs) have proven to be highly effective in extracting discriminative features from dermoscopic images, enabling improved classification performance in dermatological applications [1,2]. Benchmark studies conducted on widely used datasets such as HAM10000 and ISIC have demonstrated the strong capability of CNN-based models in achieving high diagnostic accuracy [3]. Furthermore, recent advancements have introduced hybrid architectures that combine CNNs with

transformer-based mechanisms to better capture global contextual information and enhance class discrimination [4,5]. These approaches leverage attention mechanisms to improve feature representation and model robustness [6]. Despite these advancements, most existing methods still treat skin disease diagnosis as a purely image-based classification problem, relying solely on visual information without incorporating additional clinical context and ignoring clinically relevant contextual information such as patient demographics and lesion location.

Several studies have addressed this limitation by introducing multimodal approaches that integrate dermoscopic images with structured clinical metadata, leading to improved diagnostic performance over image-only models [7]. Dermatology clinicians utilize visual examination coupled with patient-specific clinical metadata such as sex, age, lesion location, and acquisition context to create a diagnosis that is as accurate as possible. Omitting this complementary information introduces the potential for ambiguous diagnostic output, especially among lesions that may visually appear similar to each other. Efforts to mitigate this issue have led to the development of multimodal learning frameworks that fuse dermoscopic images with structured clinical metadata [8,9]. Multimodal learning frameworks have been established to provide enhanced diagnostic accuracy compared with previous methodologies; however, they currently use static fusion of images and clinical metadata, assuming that all sample images have equal relevance of both modalities to the final diagnosis.

Unfortunately, while rigid fusion models have advantages, they also have serious disadvantages. The first is that the quality and relevance of clinical metadata vary across different patient populations, and when combined uniformly, this may introduce noise into the data. The second is that static models do not have the ability to adapt to case-specific needs or to dynamically select which modality provides the most value for a particular case. Another major problem with existing multimodal models is that, when trained on imbalanced datasets, they are highly prone to overfitting and modality dominance, limiting the model's ability to generalize and, therefore, to have reliable clinical application [10,11].

DermaFusionNet is our novel multimodal skin disease classification framework that overcomes the limitations of previous work by using a gating cross-modal fusion mechanism to integrate dermoscopic images and clinical metadata in an adaptive manner. DermaFusionNet's functionality differs from conventional methodologies in that it uses learnable gating to allow for the dynamic adjustment of each modality's contribution to classification on a per-sample basis. It also introduces an auxiliary modal regularization strategy through modality-specific prediction heads to support multimodal learning while decreasing the likelihood of overfitting and preserving discriminative representations in both visual and metadata streams.

Using a comprehensive experimental methodology, the proposed architecture has been evaluated using the HAM10000 dataset. Extensive testing has shown that the proposed architecture exceeds the performance of both naively fused and unimodal baseline systems. For the proposed model, the overall classification accuracy is 89.22%, precision is 86.32%, recall is 89.22%, and F1-score is 89.08%. In addition, these evaluations confirm that gating the different modalities improves diagnostic ability, particularly for minority and visually ambiguous categories.

Materials And Methods

Related work

Due to the increasing prevalence of melanoma and other types of skin cancer and the lack of qualified dermatologists available to treat these patients, there has been a major increase in research in this field regarding the development of automated methods to classify skin diseases based on dermoscopic images. Machine learning and deep learning techniques have made substantial advancements over the last several years. Nevertheless, developing an algorithm that results in clinically useful, easily generalizable, and easily interpretable diagnosis of skin diseases continues to be a significant challenge.

CNN-Based Dermoscopic Image Classification

The initial focus of deep learning research has centered on the application of CNNs for extracting discriminative visual features from dermoscopic images. Early studies have demonstrated that several deep CNN architectures, including ResNet and DenseNet, achieve strong performance in melanoma detection and multi-class skin lesion classification tasks [12-14]. In addition, other CNN-based approaches and architectural variants have further improved classification performance through enhanced feature representation and optimized training strategies [15,16]. These models are effective in learning hierarchical spatial representations, enabling robust analysis of complex dermatological patterns.

Subsequent research has explored the use of advanced CNN-based techniques, including improved feature engineering strategies and optimized training methodologies, to enhance the reliability and robustness of skin lesion classification systems. Several studies have demonstrated improvements in CNN-based skin lesion classification through advancements in model architectures and training methodologies [17]. In addition, further works have validated the effectiveness of CNN-based approaches through large-scale evaluations and standardized benchmarking protocols conducted under controlled experimental settings [18-20]. Table *1* depicts the comparison of representative skin disease classification methods.

Method	Year	Modality	Fusion Strategy	Dataset	Key Limitation
Grant et al. [1]	2022	Image	None	HAM10000	Image-only learning
Naeem et al. [4] (SCDNet)	2022	Image	Attention based	ISIC	High complexity
Tahir et al. [5] (DSCC-Net)	2023	Image	Multi-scale CNN	HAM10000	Limited metadata use
Fraiwan et al. [7]	2022	Image + Meta	Early fusion	HAM10000	Static fusion
Saleh et al. [9]	2024	Image + Meta	Feature fusion	ISIC	No adaptive weighting
Naeem and Anees [21] (DVFNet)	2024	Image	Dual-view fusion	ISIC	Unimodal design
Saha et al. [22]	2024	Image	Detection based	HAM10000	Poor minority recall
Groh et al. [23]	2024	Image	XAI based	ISIC	No multimodal support

TABLE 1: Comparison of representative skin disease classification methods

CNN, Convolutional Neural Network; XAI, Explainable Artificial Intelligence; SCDNet, Skin Cancer Detection classifier Network; DSCC-Net, Deep learning-based Skin Cancer Classification Network; DVFNet, Dual-View Fusion Network

Though there have been advances in this area, image-based CNN models do not have a strong contextual understanding and also have difficulty distinguishing between visually similar lesion types. This problem is particularly pronounced with imbalanced datasets like HAM10000 [24,25]. The results from these studies clearly demonstrate that CNNs need

to develop more abstract representations as opposed to just local texture patterns.

Hybrid CNN-Transformer and Attention-Driven Architectures

To overcome the limitations of traditional CNNs, which are only able to capture local spatial context, further investigations have been conducted into the development of hybrid architectures. These new models use convolutional backbone networks with transformers or attention methods to provide better global context modeling, as well as greater attention to tumor boundaries [6,26].

SCDNet and DSCC-Net incorporate dense supervision and multi-scale attention mechanisms to enhance feature representation and improve class separability [4,5]. Similarly, architectures such as SNC-Net and DVFNet introduce advanced strategies like dual-view feature learning to further strengthen discriminative capability in skin lesion classification tasks [21,27]. Additional studies have proposed progressive feature refinement techniques and structured prediction frameworks to improve model performance and robustness [28,29]. These approaches emphasize more effective feature learning and prediction strategies for complex dermatological patterns [30].

These hybrid models have been shown to outperform both standalone image-based models, but they are also computationally heavy and still have a single mode to represent the data. Therefore, the incorporation of this metadata could provide additional support to a dermatologist's practice.

Multimodal Learning With Clinical Metadata

To address this, many researchers have produced hybrid diagnostic tools using clinical metadata in addition to images. Many studies to date have demonstrated improved diagnostics through the combination of image features and structured clinical data [7,31].

Recent studies on metadata-aware learning and decision support frameworks have demonstrated improvements in model robustness and a reduction in bias when clinical information is incorporated alongside visual data [8,32]. Furthermore, advanced multimodal systems have explored both feature-level and decision-level fusion strategies to enhance classification performance [9,33]. Additional works have extended these approaches by proposing more sophisticated fusion mechanisms and integration strategies for multimodal learning [34,35].

Existing multimodal systems primarily use static fusion strategies to combine visual and clinical modalities, as they assume equal weighting for both modalities on every sample. This static approach can exacerbate parasitic noise when some metadata are missing or not useful enough to inform system decisions. Therefore, existing multimodal systems are not very adaptable [10,11].

Efficiency-Oriented and Clinically Scalable Models

Lightweight and efficient architectures have been developed for ease of real-world integration, with grid-based CNNs, YOLO detection frameworks, and compact hybrid models aiming to achieve a balance between computational cost and acceptable levels of accuracy [22,36,37].

Data characteristics, population variability, and diagnostic consistency across demographic groups were the focus of other analyses [38-40]. While improved scalability can be achieved with the application of the previously mentioned models, improvements to minority class performance along with adaptive multimodal reasoning are often omitted from these systems.

Explainability and Clinical Trustworthiness

The rise of explainable AI has allowed clinicians to trust AI technology and adopt it into their practices more easily. Many researchers have worked together to produce visually interpretable evidence for the diagnostic regions that are most relevant to their work, so as to show the clinician how the AI system identifies the aforementioned regions and ultimately build up clinician trust in using AI [23,41].

Several of the studies performed to date focused on evaluating the reliability of AI-based diagnostics, analyzing how AI-based diagnostics affect patient outcomes, and developing new methods that take demographic information into consideration to determine clinical efficacy [42,43]. However, most of the current explainable models are based on a single input and do not allow clinicians to interpret multiple types modalities of information in multimodal.

Comparative Analysis of Existing Methods

Table *1* depicts the different approaches to skin disease classification by demonstrating how each method uses modalities, how the modalities are fused to achieve the best results, which datasets were used to develop the models, and the major weaknesses of each approach. The findings of our analysis conclude that existing skin disease classification approaches have either relied solely on visual information or fused modalities using only static methods, without the capability to adaptively weigh the modalities or use auxiliary regularization strategies.

Research Gap and Motivation

The preceding analysis identifies several important limitations in current skin disease classification research. Although prior studies have explored image-based, hybrid, and multimodal diagnostic systems, significant challenges remain in three key areas: (1) limited and inconsistent integration of clinically relevant metadata alongside dermoscopic imagery, (2) insufficient exploration of adaptive sample-specific multimodal fusion strategies that dynamically adjust modality contributions based on case context, and (3) limited investigation of auxiliary regularization mechanisms to stabilize multimodal learning and reduce modality dominance under imbalanced clinical settings [44,45].

These unresolved challenges motivate the development of DermaFusionNet, which introduces a unified framework combining gated cross-modal fusion with auxiliary modal regularization to improve robustness, interpretability, and clinically relevant diagnostic performance.

Dataset and preprocessing

High-quality labeled datasets along with thoroughly defined preprocessing pipelines are necessary for reliably assessing an automatic skin disease classification model's performance. In the present research, we used a dermoscopic benchmark dataset with high acceptance levels and applied standardized preprocessing procedures to ensure that the dataset was consistent, diagnostic related, and could provide for equal comparisons among all experimental contexts. This section presents a thorough characterization of the dataset, including specifications as to the type of class imbalance, distribution of classes, and image quality rating prior to implementing any models.

HAM10000 Dataset Overview

The HAM10000 dataset [46], publicly available at https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000, consists of thousands of skin lesion photographs that can be used to develop automated systems for diagnosing skin lesions. This dataset contains 10,015 dermoscopy images collected from a variety of locations and settings. It includes images that represent real variations in the diagnosis of skin conditions. HAM10000 contains images of seven different types of skin lesions classified as either benign or malignant, thus allowing for the classification of multiple types of skin diseases.

The HAM10000 dataset contains visual images provided with expert-verified diagnostic labels based on histopathology, follow-up examinations, or consensus evaluations. The dataset also includes clinically relevant metadata associated with the image files, such as the patient's age, sex, lesion localization, and source of image acquisition. This combination of visual and clinical information makes HAM10000 an excellent resource for developing and evaluating multimodal diagnostic systems for skin diseases. With its broad representation across multiple skin conditions, consistent labeling system, and widespread use in the research literature, HAM10000 provides a robust benchmark for measuring the performance and generalizability of skin disease classification models.

Class Distribution Analysis

The analysis of class distribution shows a class imbalance throughout the HAM10000 dataset [46], which is representative of most real-world medical datasets. Lesions classified as melanocytic nevi greatly outnumber the other classes, resulting in a highly imbalanced dataset, while tumors and other malignant or rare lesions are present in dramatically lower numbers. Many automated classification systems face significant challenges in addressing this issue, since the models tend to become biased toward high-volume classes.

Table *2* summarizes the class-wise distribution of images across the seven lesion categories. As observed, the dataset is dominated by benign nevi samples, whereas dermatofibroma and vascular lesions constitute a very small fraction of the dataset. This imbalance necessitates careful experimental design and evaluation using robust metrics beyond simple accuracy.

Lesion Class	Number of Images
Melanocytic Nevi (nv)	6,705
Melanoma (mel)	1,113
Benign Keratosis-like Lesions (bkl)	1,099
Basal Cell Carcinoma (bcc)	514
Actinic Keratoses (akiec)	327
Vascular Lesions (vasc)	142
Dermatofibroma (df)	115
Total	10,015

TABLE 2: Class-wise distribution of the HAM10000 dataset

Visual Inspection and Data Quality

Before preparing data for model training, we performed a visual inspection of the dermoscopic image dataset for both the quality of the data and to assess the variability between different classes of dermoscopic images taken by the same provider. To identify potential problems that may impact the algorithm performance with respect to the dermoscopic image dataset, we selected a small subset of representative images from each lesion type examined during our inspection and analyzed for potential discrepancies. Some of the issues identified during this process included low contrast, inconsistency in color across the same image, variability in the amount of illumination, presence of hair or other artifacts, and background noise in the images.

It was noted by visual inspection that there was a wide variety of characteristics from class to class as well as within each class. The visual features of benign versus malignant lesions can often have the same visual appearance, making it difficult for an automated system to discriminate between them. Obtaining images from different acquisition devices, protocols, and settings would also lead to differences in the resolution, illumination, and scale of the photos. Therefore, effective preprocessing and feature extraction methods will be critical in accommodating for the variety of conditions found in the real world.

Representative dermoscopic samples from all seven classes are illustrated in Figure 1, highlighting the diversity of lesion appearance and the inherent diagnostic complexity of the task.

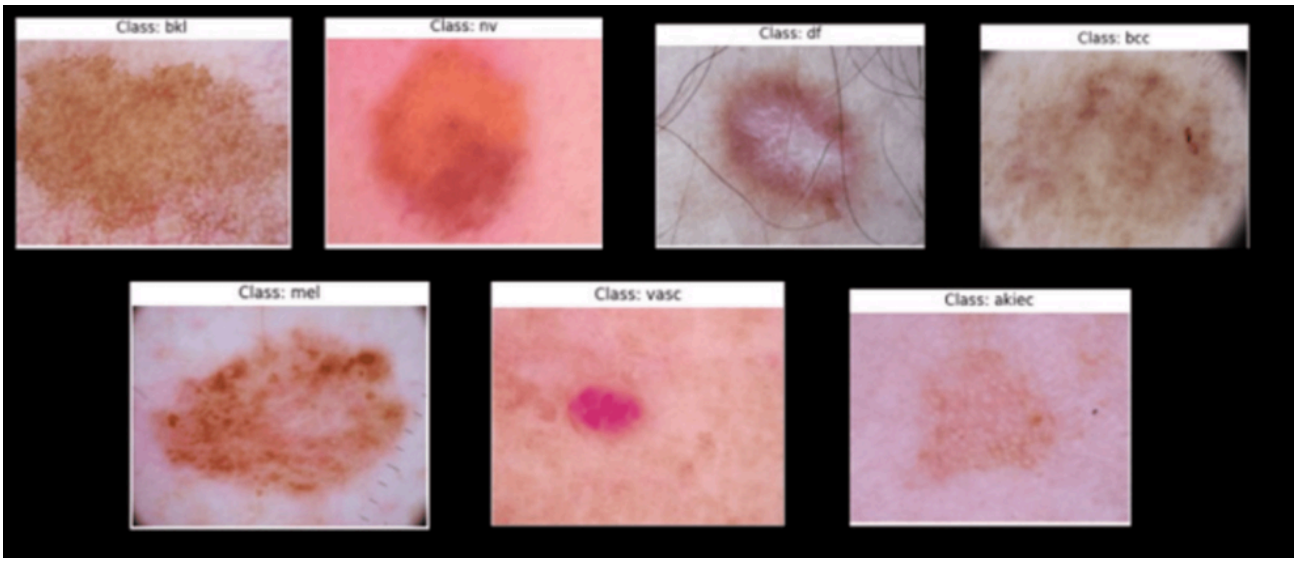

FIGURE 1: Representative dermoscopic images from the HAM10000 dataset illustrating the visual diversity across the seven lesion categories, including melanocytic nevi (nv), melanoma (mel), benign keratosis-like lesions (bkl), basal cell carcinoma (bcc), actinic keratoses (akiec), vascular lesions (vasc), and dermatofibroma (df). The samples highlight substantial intra-class variability and inter-class visual similarity, underscoring the complexity of automated skin disease classification

Image Preprocessing

The HAM10000 dataset contains dermoscopic photographs that were acquired from a variety of clinical settings and imaging devices, creating significant differences in illumination, contrast, color saturation, and background conditions. This variability in data capture can introduce extraneous noise and therefore hurt the ability of a model to generalize, since these spurious correlations are not related to normal lesion morphology. To combat this issue, a preprocessing standard is used to preprocess the images so that they appear visually similar, but retain diagnostically useful features such as pigment network features, streaks, globules, and irregular border features. The preprocessing approach taken in this work is intentionally designed to be very lightweight, reproducible, and clinically conservative; it focuses on ensuring the same resolution and contrast as will be found in real-world clinical use cases but not performing excessive and aggressive removal of artefacts that could negatively impact the visual integrity of most lesions' borders.

Preprocessing objectives: The preprocessing pipeline targets the following objectives:

• Geometric Standardization: enforce a consistent input resolution for efficient batch training and stable feature extraction.

• Photometric Normalization: reduce illumination and contrast variability while retaining local lesion texture.

• Color-Space Stabilization: apply enhancement primarily on luminance to avoid color shifts that may confound diagnosis.

• Storage Efficiency: save preprocessed images in a compact format without noticeable diagnostic loss.

Algorithm: Standardized resize + luminance-CLAHE enhancement: Algorithm 1 summarizes the image preprocessing steps used in this study. The method operates in the LAB color space to isolate luminance (L) from chrominance (a,b). Contrast enhancement is applied only to the luminance channel using CLAHE (Contrast Limited Adaptive Histogram

Equalization), which improves the visibility of subtle lesion patterns under uneven illumination. The contrast-limiting mechanism prevents over-amplification of noise by clipping the histogram at a configurable threshold. After enhancement, channels are merged and converted back to RGB/BGR for downstream processing.

The CLAHE operation applied to the luminance channel can be expressed as:

$L' = \text{CLAHE}(L, \lambda, \tau)$,

where L represents the luminance channel, λ denotes the clip limit, and τ represents the tile grid size used for local histogram equalization.

The numerical metadata is normalized using z-score normalization as:

$A_k = (A_k - \mu_k)/(\sigma_k + \varepsilon)$

where μ_k and σ_k represent the mean and standard deviation of the attribute A_k, and ε is a small constant to avoid division by zero.

When working with the LAB color space, luminance can be separated from the color properties (a, b). This enables the application of contrast-enhancement techniques, such as CLAHE, only to the luminance channel, thereby enhancing the visualization of very opaque lesions affected by inhomogeneous lighting. The contrast-limiting technique allows noise to be clipped at a predefined value before being amplified beyond the specified limit. Once the image has been enhanced, the separate channels can be combined into a single image for processing in RGB format.

Algorithm 1: Image Preprocessing Pipeline (Resize + LAB-CLAHE)

Require: Input dermoscopic image I (RGB/BGR), target size S × S, CLAHE clip limit λ, tile size τ × τ

Ensure: Preprocessed image $\hat{I}$ of size S × S

1: Resize: Ir ← Resize (I, S, S) using area interpolation

2: Color transform: (L, a, b) ← RGB2LAB(Ir)

3: CLAHE on luminance:

4: L′ ← CLAHE(L; λ, τ)

5: Merge channels: Ilab ← Merge(L′, a, b)

6: Inverse transform: $\hat{I}$ ← LAB2RGB(Ilab)

7: Output: return $\hat{I} = 0$

Technical rationale and design choices: In this subsection, we discussed preprocess steps in depth and technical manner.

1) Resize to a fixed spatial resolution: To make sure that all of the pictures match the typical input size for the ImageNet pre-trained ConvNeXt backbones, they were reduced to 224 × 224 pixels. In order to create less aliasing artifacts when enlarging or reducing the size of an image, the resizing process uses area interpolation. In addition, using this technique produces a more consistent preservation of lesion boundaries than nearest-neighbor techniques. Additionally, by preserving stable convergence in the training phase and reducing operational costs during training, this method maximizes operational efficiency.

2) CLAHE in LAB space to preserve diagnostic color cues: Clinically meaningful differences of color may be affected by direct RGB histogram equalization, as it distorts chromatic information. To alleviate this problem, images are converted to LAB space where only the luminance channel is enhanced during preprocessing. Enhancement enhances local contrast while preserving the chromaticity (a,b) close to where it was prior to preprocessing, thus preserving clinically relevant color distributions.

3) Contrast-limited enhancement to avoid noise amplification: The adaptation of Classic Adaptive Histogram Equalization has been known to exaggerate noise from sensors, such as hairs detected through imaging, and background textures created by lighting. Meanwhile, CLAHE seeks to address this problem by limiting the scaling of a histogram to a maximum clipping limit. In other words, this prevents excessive contrast stretching of homogeneous regions of an image during histogram equalization due to CLHA system limitations. For instance, in CLHA implementations at our organization, we set our clipping limit to be approximately two times $\lambda = 2.0$ and created a grid of tile areas measuring 8 $\times$ 8 regions. This was determined to be a reasonable balance between increasing image visibility and minimizing artifacts produced through these methods.

4) Minimalist preprocessing to retain morphological integrity: Aggressive denoising, smoothing, or hair-removal operations were not applied in these studies. Though some cosmetic image enhancement techniques can enhance the aesthetic quality of images, they may also suppress thin and intricate features that are diagnostic indicators. Consequently, to protect morphology and allow the DermaFusionNet to develop reliably feature detecting models that accurately correlate to patient anatomy, a pipeline of preservation and replicability has been established.

Output generation and storage: JPEG compressed with a quality factor of 92 was used, thus saving space without incurring noticeable compression artefacts. The final output of this stage is arranged into directories based on both the train and test splits, along with their respective class labels. Therefore, the training data can be easily reproduced and audited. Throughout the preprocessing step, we generated consistent and contrast-normalized dermoscopic images, allowing for the utilization of ConvNeXt features extracted using DermaFusionNet.

Metadata Encoding

When making a diagnosis in dermatology practice, clinicians do not rely solely on visual examination; they also take into account patient-specific context. Age, gender, lesion location, source of acquisition, and many other attributes unique to each patient provide context that can be used to differentiate between visually similar skin lesions. DermaFusionNet uses an external metadata encoding pipeline to transform clinical metadata into numerical representations that can be utilized by deep learning algorithms.

Clinical metadata description: For each dermoscopic image, we used the following metadata attributes, if they were available: Source Dataset [acquisition origin (HAM10000)], Age (patient age in years), Sex (male or female), and Lesion Localization [anatomical site such as trunk, scalp, and extremities]. The metadata attribute types include both categorical and numerical values and therefore require both categorical and numerical encoding methods to preserve the semantics and numerical stability during training times.

Algorithm: Hybrid categorical-numerical metadata encoding: As per Algorithm 2, this paper explores the metadata preprocessing and encoding. Two separate techniques were used to process and encode metadata, one for each type of metadata that exists. Categorical variables were encoded into one-hot vectors, and the numerical metadata were normalized to produce a set of standard deviations, thereby allowing concatenation of the two different vector types to produce a common fixed-length embedding for all the metadata.

Algorithm 2: Metadata Encoding Pipeline

Require: Metadata table M with N samples and attributes $\{A_1, A_2, \ldots, A_K\}$.

Ensure: Encoded metadata matrix $X_{meta} \in R^{N \times D}$

1: for each attribute A_k in M do

2: if A_k is categorical then

3: Construct vocabulary V_k of unique categories

4: Map each category to an index

5: Encode A_k using one-hot representation

6: else

7: Replace missing values with zero

8: Standardize using z-score normalization:

$$A_k = (A_k - \mu_k)/(\sigma_k + \epsilon)$$

9: end if

10: Append encoded representation to feature list

11: end for

12: Concatenate all encoded features to obtain X_{meta} can be expressed as:

F_{meta} = concat (A_1, A_2, . . . , A_K),

13: Output: return X_{meta} = 0

Technical rationale and design choices: In this subsection we discussed metadata encoding steps in depth and technical manner.

1) One-hot encoding for categorical stability: One hot encoding of clinically categorical data is done without ordinal relationships between categories. By using one hot encoding, we preserve independence of the categories and prevent any numerical ordering from being introduced into the model that may lead to spurious results. This will also maintain the ease of interpretation; the model will output based on a known clinical category, which means there is no order associated with the various one hot encoded dimensions of the categorical clinical variables.

2) Z-score normalization for numerical consistency: As a continuous numerical attribute, the age of a patient is normalized with Z-score standardization. Normalizing the age results in having a mean of zero and a variance equal to 1. The Z-score standardization prevents the age-related attributes from dominating the gradient updates and thus ensures that the algorithm converges to an optimal solution in a stable manner.

3) Robust handling of missing metadata: Clinical data often contains many gaps in the metadata. In order for the information to be reliable, we have developed a way to address missing data through two specific ways: First, we have created an additional category called "unknown" for categorical attribute types in the dataset where a value is missing; we call this method categorical missing value representation. Second, we use a method called "zero imputation" by replacing the empty value for numerical attributes with a zero value. Both methods allow the algorithm to still account for samples without discarding them from the model. Therefore, our models are able to learn from both uncertain and confident representations at the same time.

4) Unified metadata embedding: In the case of the HAM10000 dataset, the encoding of these object types has a combined encoding of dimensionality D = 21. The combination of these various types of encodings creates one large matrix that provides for fast learning of how to correlate the images of these two object types. Then, at the end of the learning phase, we can project the combined encoding vector into the visual feature space via the use of a learnable metadata encoder, allowing us to make efficient use of the connections established between the image and metadata.

Role in multimodal learning: DermaFusionNet receives the vector of encoded metadata in conjunction with the visual modality as auxiliary inputs to DermaFusionNet. Instead of concatenating the metadata with the visual feature vectors directly, the proposed approach transforms the vector of metadata to a latent representation with the same dimensionality as the image feature vector so that it can be used in a gated cross-modal fusion architecture. This architecture allows for the network to dynamically determine how much clinical context is to be included into its predictions based on how relevant it is to the particular sample.

Data Splitting Strategy

It is very important to have a strong and fair method for dividing data into three different parts (train, validation, and test) so that we can fairly evaluate our models. This is especially true for model evaluation when using machine learning in medical images, because there are frequent problems with class imbalances and bias in medical image datasets. This paper describes how the HAM10000 dataset was split into three different subsets, training, validation, and testing, by using a stratified sampling technique to maintain the class proportions as found in the HAM10000 dataset on all three subsets that were created.

The datasets included are organized into training, validation, and testing sets, where 70% is a training set, 15% is a validation set, and 15% is a testing set. To ensure that the individual classes of diseases that have a low incidence (df) and vascular lesions (vasc) are fairly represented in each of the three sets, stratification was performed for the lesion class labels. This strategy protects against misrepresentations of the model's performance.

The HAM10000 dataset does not have unique patient IDs provided for each sample like some patient-splitting methods. Therefore, it is necessary to use lesion-level stratifications, which is consistent with the protocols used by many other previous studies regarding the classification and processing of skin lesions. This provides a balance between the need for representative evaluations without artificially inflating performance scores, while also allowing practical limitations to be accommodated.

The distribution of classes within the training, validation, and testing splits is outlined in Table *3*, and we can see that with respect to all three sets, the nv class is the most significant class in these datasets, which has remained in proportion to all other classes; however, smaller minority classes provide enough number of examples so that they can be adequately trained and evaluated.

Class	Train	Validation	Test	Total
akiec	229	49	49	327
bcc	360	77	77	514
bkl	769	165	165	1,099
df	81	17	17	115
mel	779	167	167	1,113
nv	4,693	1,006	1,006	6,705
vasc	99	21	22	142
Total	7,010	1,502	1,503	10,015

TABLE 3: Class-wise distribution of the HAM10000 dataset across training, validation, and test splits

Rationale and Impact on Model Training

In multimodal learning situations where the model has to learn the imagery as well as the clinical data, the data distribution must remain the same for all levels of the model and not be overfit to the most common categories. It is very important to maintain the same distribution of the different classes across these two phases. The stratification of the

splits allows for an accurate representation of both image feature distributions and other metadata between the training and evaluation phases, allowing for accurate estimation of the generalization performance of the model.

Lastly, there is no application of oversampling or synthetic augmenting at the dataset level when it is split into training and test datasets. Instead of using oversampling augmentation to balance the distribution of classes in the split dataset, we will balance the distribution of the training samples during model training using a weighted sampling scheme. This approach enables the model to effectively learn from the naturally distributed dataset and minimizes the bias toward the majority classes. The approach also provides a realistic simulation of the clinical deployment of the model and avoids over-optimistic predictions of performance.

Overall, the methodology for splitting data into training and test datasets has established the framework to conduct all experiments in a way that allows for all subsequent experiments to be repeatable, equitable, and comparable with previous ablation studies and base models.

Proposed method

DermaFusionNet is a multimodal deep learning system that combines dermoscopic visual information and structured clinical metadata to classify skin diseases accurately. This architecture includes a powerful convolutional model along with an encoder for learning metadata, and a mechanism for merging or fusing the information across both modes using a learnable process. In addition, an auxiliary regularization technique applied during training increases the robustness of each modal and ensures that one modal does not dominate the classification, thereby creating a model that is less susceptible to variations in one of the modalities.

An overview of the proposed architecture is illustrated in Figure 2, followed by detailed descriptions of each component.

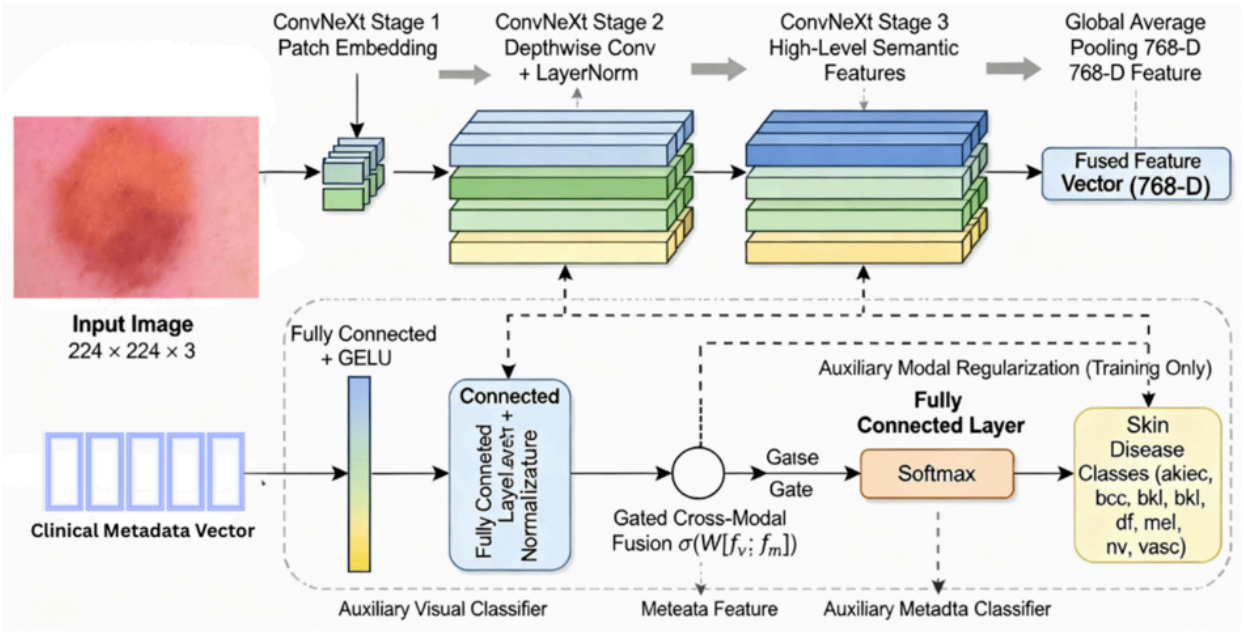

FIGURE 2: Overall architecture of DermaFusionNet. The model consists of a ConvNeXt-based visual encoder, a clinical metadata encoder, a gated cross-modal fusion module, and a classification head. Auxiliary modal regularization is applied during training to improve robustness against missing or noisy modalities

Overall Architecture Overview

DermaFusionNet follows a dual-stream design consisting of a visual branch and a metadata branch, which are fused at an intermediate representation level using a gated cross-modal fusion module. Let $\mathbf{x}_v \in \mathbb{R}^{H\times W\times 3}$ denote an input dermoscopic image and $\mathbf{x}_m \in \mathbb{R}^{d_m}$ represent the corresponding encoded clinical metadata vector.

The visual stream extracts high-level semantic features $\mathbf{f}_v \in \mathbb{R}^{768}$ using a ConvNeXt-based backbone, while the metadata stream transforms heterogeneous clinical attributes into a compact embedding $\mathbf{f}_m \in \mathbb{R}^{768}$. These modality-specific representations are then adaptively fused to produce a unified feature vector $\mathbf{f}_{\text{fusion}}$, which is subsequently passed to a classification head for lesion category prediction.

DermaFusionNet provides a modular architecture that enables a more flexible system with respect to any modality that may be available and/or needed at the time.

Visual Feature Extractor

Through the use of ablation studies, we investigate how effectively the visual feature extractor in DermaFusionNet extracts meaningful representation from dermoscopic images. In our implementation, we utilize ConvNeXt-Small, a CNN architecture that has demonstrated state-of-the-art capabilities in multiple domains including medical image analysis by leveraging its large number of learned weights to generalize better across many different datasets than other architectures have done.

Given an input image $\mathbf{x}_v$, the backbone produces a feature map $\mathbf{F}_v \in \mathbb{R}^{C\times h\times w}$, which is spatially aggregated using global average pooling:

$$\mathbf{f}_v = \mathrm{GAP}(\mathbf{F}_v) \tag{1}$$

where $\mathbf{f}_v \in \mathbb{R}^{768}$ represents embedding that captures all texture, color distribution, and morphology of the lesion. These characteristics are critical in distinguishing between skin conditions that may appear similar visually.

Gated Cross-Modal Fusion Module

To combine visual features with metadata features, DermaFusionNet introduces a gated cross-modal fusion technique, which adaptively assigns weights to each modality, using their combined contexts. Rather than simply concatenating the two modalities together, the gating strategy allows for dynamic highlighting or suppression of the contribution of each modality depending upon its relevance at that moment.

The first step is to concatenate both visual embedding and metadata embeddings together:

$$\mathbf{z} = [\mathbf{f}_v; \mathbf{f}_m] \tag{2}$$

where $[\cdot;\cdot]$ denotes vector concatenation. A gating vector $\mathbf{g} \in \mathbb{R}^{768}$ is then computed as:

$$\mathbf{g} = \sigma(\mathbf{W}_g\mathbf{z} + \mathbf{b}_g) \tag{3}$$

with $\sigma(\cdot)$ denotes the sigmoid activation function. The final fused representation is obtained as:

$$\mathbf{f}_{\text{fusion}} = \mathbf{g} \odot \mathbf{f}_v + (1 - \mathbf{g}) \odot \mathbf{f}_m \tag{4}$$

where $\odot$ represents element-wise multiplication. Element-wise multiplication allows for fine control of how the modalities interact and provides for high resolution and fine granularity in describing those effects, which becomes especially advantageous when there is variability in the quality of metadata associated with each sample.

Auxiliary Modal Regularization

In order to improve the robustness of the model against over-weighting one modality, DermaFusionNet has been developed to use auxiliary modal regularization during the training process. This auxiliary modal regularization consists of classification heads that can be added to the embeddings of each individual modality to encourage those branches to learn sufficiently meaningful representations independently of one another.

Let $\mathcal{L}_v$ and $\mathcal{L}_m$ denote auxiliary classification losses computed from $\mathbf{f}_v$ and $\mathbf{f}_m$, respectively. These losses act as regularizers that stabilize multimodal learning and improve generalization under modality degradation scenarios.

The removal of auxiliary regularization in Ablation A1 demonstrates, through empirical evidence, a significant drop in performance due to the lack of auxiliary input.

Loss Function and Optimization

The main goal of DermaFusionNet is to reduce the categorical cross-entropy loss associated with skin disease classification, which is calculated through the use of the ground-truth label(s) *y* and the predicted class probabilities $\hat{\boldsymbol{y}}$, making the overall loss computed as follows:

$$\mathcal{L}_{\text{main}} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \tag{5}$$

where *C* denotes the number of lesion classes. When auxiliary regularization is enabled, the total training loss becomes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \lambda(\mathcal{L}_v + \mathcal{L}_m) \tag{6}$$

with $\boldsymbol{\lambda}$ controlling the contribution of auxiliary losses.

The method used to optimize networks will be AdamW, which includes a cosine annealing learning rate schedule. In addition, gradient clipping will be used. AdamW with cosine annealing and gradient clipping allows the parameters to converge rapidly and provides good generalization capabilities across all presented ablation configurations.

Experimental setup

This section describes the methods in which the DermaFusionNet framework was tested and how it was evaluated. We discuss the training configuration, the hyperparameters used, the evaluation metrics for overall performance, and how the ablation studies were designed to evaluate each architectural component of DermaFusionNet. All of these experiments are set up to provide a fair comparison, ensure reproducibility, and provide clinical value.

Training Configuration and Hyperparameters

All of the experiments in this paper used the PyTorch deep learning framework and were conducted in an environment that supports GPU execution. The backbone of DermaFusionNet is built upon the pre-trained ConvNeXt-Small weights from ImageNet, enabling rapid convergence and better generalization of features extracted from dermoscopic images. The metadata encoder and metadata fusion modules were trained completely from scratch.

The images we input into our model have a consistent size, which is 224 × 224, and we normalize those images according to the statistics of the ImageNet dataset. As part of the data augmentation process during training, we randomly crop each image within the 224 × 224 pixel area, horizontally flip each image, and occasionally apply minor color adjustments. All of these augmentation strategies are designed to improve the robustness of the model to variations in appearance that can be observed in dermoscopy images.

By utilizing a cosine annealing LR scheduler, we can effectively regulate the LR as it decreases over the course of training.

The AdamW optimization algorithm was initially set at 2×10^{-4} for our beginning LR and 1×10^{-4} for our weight decay value. Depending upon the experiment conducted, the model underwent 30 epochs of training prior to the selection of an appropriate model according to validation performance. In order to reduce the effects of class imbalance on model performance metrics and ultimately increase accuracy on minority represented lesions, a class weight sampling strategy was implemented. This allowed us to maintain natural distributions of our training sample but ensured each class received an equivalent amount of training attention.

Evaluation Metrics

Assessing the classification performance of a model for data with class imbalance requires several evaluation metrics. Total accuracy considers all classifications as correct without accounting for how many different classes there may be in relation to each other. Balanced accuracy compensates for this by calculating the average of the recall across all classes, therefore giving equal weight to the performance of the model across all different classifications irrespective of how large or small their respective class size might be. The evaluation metrics are formally defined as follows:

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN)$$

$$\text{Precision} = TP/(TP + FP)$$

$$\text{Recall} = TP/(TP + FN)$$

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

$$\text{Balanced Accuracy} = (1/C) \times \Sigma\, \text{Recall}_c$$

In the final analysis, both macro- and weighted precision, recall, and F1-score were computed to determine how well the model was able to discriminate between classes, as well as how well the model as a whole is capable of making reliable predictions. To assess the reproducibility of predictions over and above mere chance, the Matthews correlation coefficient (MCC) and Cohen's kappa coefficient have also been included in the report. Collectively, these metrics provide a comprehensive, meaningful, and clinically useful assessment of the model's performance on both frequently occurring and infrequently occurring lesion classes.

Ablation Study Design

In order to perform a systematic analysis of the individual contributions of each architectural component in DermaFusionNet, we will conduct a set of controlled ablation studies, where each variant is tested on the same data splits, under the same training configuration, using the same optimization strategy and evaluation protocol as the complete DermaFusionNet model, so that performance differences observed can be directly attributed to architectural changes and not to experimental bias.

Ablation-1: Without Auxiliary Modal Regularization: In setting A1, during training, we remove the auxiliary modal regularizations. In this case, the auxiliary supervision used in the individual visual and metadata branches is no longer applied. As a result, only the primary classification loss from the fused representation is used to optimize the model. The purpose of this experiment is to assess how auxiliary regularization contributes to the stabilization of multimodal learning and promotes the extraction of modality-specific discriminative features.

Ablation-2: Fusion Without Gating: The second ablation (A2) studies the benefit of this gated fusion method by using a more simplistic concatenation of the feature representations with no adaptive weighting of the modalities. In this way we can determine if the dynamic gating strategy has a significantly better performance compared to standard fusion methods when used for classifying skin diseases across multiple modalities.

Ablation-3: Image-Only Model: Styled on the same theme as the previous two ablations, the third ablation (A3) provides insight into the impact of clinical data by completely removing that data stream and therefore testing only the ConvNeXt-based visual encoder and its respective classification head. As such, A3 represents the image-only version of the model. Dummy zero-value metadata vectors were used in lieu of actual metadata to ensure consistency with the model architecture. This ablation quantifies the improvement or added value gained by including structured clinical data with dermoscopic images.

Results And Discussion

Experimental results and analysis

The results of a detailed evaluation of the proposed DermaFusionNet framework, which utilized the HAM10000 dataset, are included in this section. The first part of the evaluation provides an overview of the quantification of performance of the entire multimodal model. Then, the second area of evaluation includes detailed performance at class level, comparison of ablation studies, and confusion matrix results. Multiple methods were used to generate the evaluation results for each of the above-mentioned aspects, such as a range of metrics providing holistic and clinically valuable insights regarding the performance of the system, including comparative performance against imbalanced data classes. All the data provided in this evaluation section are based on the results generated from the held-out test set derived from the highest-performing model based on validation performance.

Main Results

In Table 4, the comprehensive performance summary for DermaFusionNet on the HAM10000 dataset is presented. As shown by this information, the proposed method is able to achieve excellent classification performance across all evaluation metrics. This indicates that gated cross-modal fusion is an effective method of integrating clinical metadata with dermoscopic images.

Metric	Result
Accuracy	0.8922
Balanced Accuracy	0.7738
Precision (Macro)	0.8632
Recall (Macro)	0.7738
F1-score (Macro)	0.8099
Precision (Weighted)	0.8927
Recall (Weighted)	0.8922
F1-score (Weighted)	0.8908
Matthews Correlation Coefficient (MCC)	0.7916
Cohen's Kappa	0.7913

TABLE 4: Overall performance of DermaFusionNet on the HAM10000 test set

Table 4 indicates that DermaFusionNet obtains an overall accuracy of 89.22% and a weighted F1-score of 89.08%, demonstrating that it predicts well for both dominant and minority lesion types. Although the balanced accuracy is 77.38%, this indicates that DermaFusionNet performs exceptionally well when considering all classes, particularly given

the fact that there are substantial imbalances between the classes in the training dataset. In addition, high macro-level precision indicates that there are very few false-positive predictions made by DermaFusionNet when predicting all types of lesions.

The consistently high MCC and Cohen's kappa metrics demonstrate that there is a strong agreement between prediction and ground-truth labels beyond chance. Therefore, the reliability of our proposed multimodal framework has been validated. Our results suggest that by integrating clinical metadata into the model using a gated cross-modal fusion approach, we are able to significantly improve diagnostic accuracy over a model that relies solely on image data. Subsequent ablation studies also demonstrated the enhanced performance of the model when clinical metadata was incorporated.

Class-Wise Performance Analysis

A class-wise performance analysis of DermaFusionNet's diagnostic performance was conducted on the HAM10000 testing dataset, allowing us to see how robustly DermaFusionNet performs across the skin lesion types included in the testing dataset. This analysis will provide both the majority and minority skin lesion types that are most likely to appear in clinical settings, thus giving us a better understanding of how DermaFusionNet could be used in a clinical setting to detect rare and potentially dangerous skin lesions.

All of the metrics presented in Table *5* (precision, recall, F1-score, and support) show that DermaFusionNet performs well overall; however, it performs particularly well in the dominant classes such as melanocytic nevi (nv), and it still performs competitively in less represented classes.

Class	Precision	Recall	F1-score	Support
akiec	0.8667	0.5306	0.6582	49
bcc	0.8684	0.8571	0.8627	77
bkl	0.7661	0.7939	0.7798	165
df	0.9231	0.7059	0.8	17
mel	0.7283	0.7545	0.7412	167
nv	0.9422	0.9563	0.9492	1,006
vasc	0.9474	0.8182	0.878	22

TABLE 5: Class-wise performance of DermaFusionNet on the HAM10000 test set

DermaFusionNet performs best on the nv class; it achieved an F1-score of 94.92%, indicating that the model was able to learn discriminative patterns successfully due to a large number of training samples available to train the model. In addition, it performed well on other clinically significant malignant lesion types, including basal cell carcinoma (bcc) and melanoma (mel), with F1-scores of 86.27% and 74.12%, respectively.

For rare classes, including dermatofibroma (df) and vascular lesions (vasc), the model has high precision values; therefore, it has low false-positive rates. The lower recall values for akiec and df, on the other hand, suggest that these classes are difficult due to their visual ambiguity and sparse representation, thus demonstrating the necessity of multimodal fusion and balanced training techniques.

A bar chart illustrating class-specific performance trends is shown in Figure *3* through F1-scores for all types of lesions. This provides an easy-to-read example of how well DermaFusionNet performed consistently across all different types of lesions. It also highlights the large differences in performance across highly imbalanced classes.

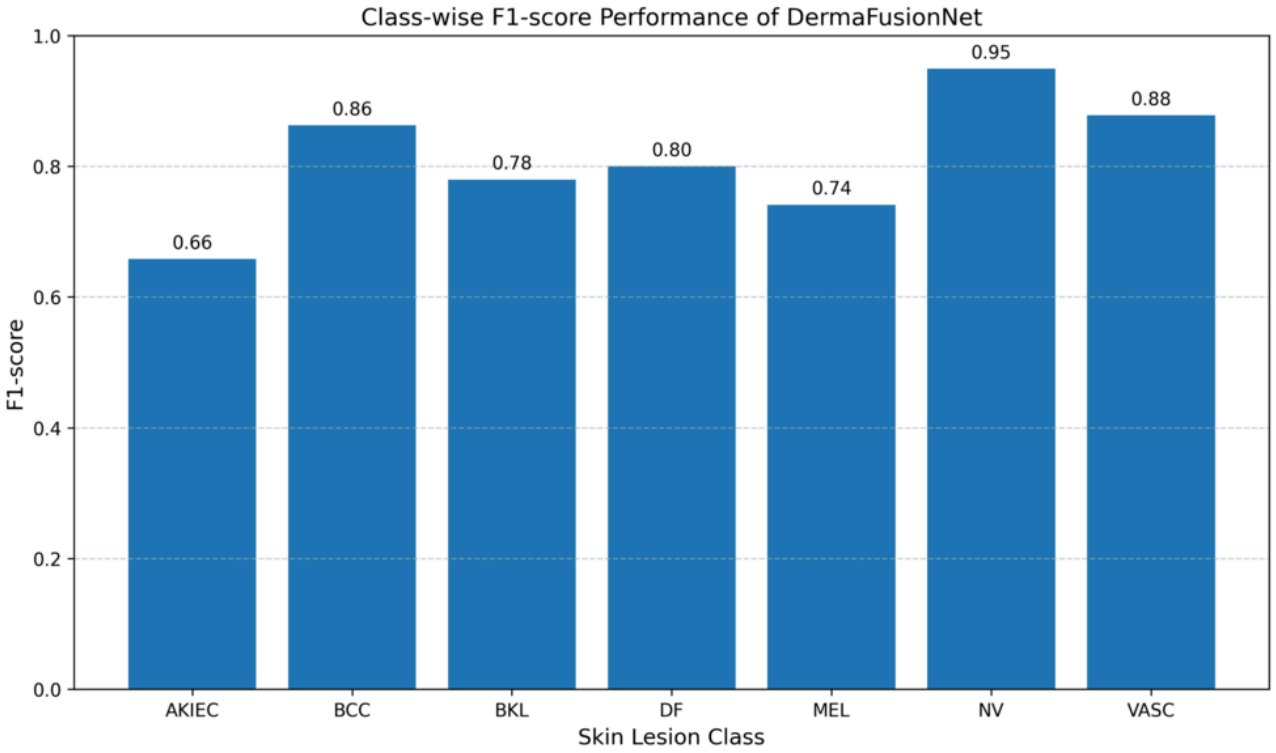

FIGURE 3: Class-wise F1-score distribution of DermaFusionNet on the HAM10000 test set

Based on the results of class-wise testing of DermaFusion-Net capabilities, it has been demonstrated that this model is capable of identifying not only the most common types of skin tumors but also successfully providing diagnostic assistance for those lesions that are less common and/or critical to treat from a clinical perspective. This indicates that DermaFusionNet is the best option for dermatological screening in actual clinical practice.

Ablation Study Results

An extensive ablation study was conducted to systematically evaluate the value that each individual architectural component contributes to DermaFusionNet. This study aimed to quantify how much auxiliary modal regularization, gated cross-modal fusion, and the use of metadata influence performance when classifying images. Four different configurations were evaluated under the same training conditions: the complete DermaFusionNet model and three ablated models (A1) no auxiliary modal regularization, (A2) fusion with no gating, and (A3) learning image data only without metadata. The models were evaluated on the HAM10000 test set using several measures to ensure overall robustness and clinical relevance.

Overall Performance Comparison

A complete quantitative comparison of the DermaFusionNet architecture and all its ablation variants is provided in Table 6. Based on nearly every evaluation metric, the full model produces superior performance consistently. Therefore, it confirms the value of jointly modeling both dermoscopic images and clinical metadata via gated fusion and auxiliary supervision.

According to Table 6, eliminating the use of auxiliary modal regularization (A1) results in a considerable decrease in both the weighted F1-score (which dropped from 0.8908 to 0.8729) and the MCC. This suggests poorer generalization ability for the model across the various imbalanced lesion classes. Thus, it can be concluded that auxiliary supervision is an efficient regularizer that prevents overfitting on the most prevalent visual patterns while also promoting the balanced acquisition of multimodal features.

Model	Accuracy	Balanced Accuracy	Precision (Macro)	Recall (Macro)	F1 (Macro)	Precision (Weighted)	Recall (Weighted)	F1 (Weighted)	MCC	Kappa
DermaFusionNet (Full)	0.8922	0.7738	0.8632	0.7738	0.8099	0.8927	0.8922	0.8908	0.7916	0.7913
A1: No Auxiliary Regularization	0.8749	0.7547	0.8634	0.7547	0.8005	0.8739	0.8749	0.8729	0.7554	0.7546
A2: Fusion Without Gating	0.8829	0.8092	0.8523	0.8092	0.829	0.8836	0.8829	0.8828	0.7754	0.7752
A3: Image-Only Model	0.8829	0.7794	0.8517	0.7794	0.8096	0.8835	0.8829	0.8818	0.7736	0.7732

TABLE 6: Overall performance comparison of DermaFusionNet and its ablation variants on the HAM10000 test set

MCC, Matthews Correlation Coefficient

While the A2 arrangement achieves relatively competitive aggregate performance through simple feature concatenation, its lack of adaptive modality weighting limits its flexibility compared with the full DermaFusionNet architecture. The gated fusion mechanism was specifically designed not only to improve benchmark metrics but to dynamically regulate modality contribution based on sample-specific clinical relevance. Although the aggregate numerical gap between gated and ungated fusion appears moderate, gated fusion provides improved robustness against modality dominance and greater adaptability when metadata quality, completeness, or diagnostic relevance vary across cases. This adaptive property is particularly valuable for real-world clinical deployment, where multimodal reliability may be more important than marginal benchmark improvements alone.

The importance of including metadata information in an image-only model (A3) has also been clearly illustrated by this study. While the competitive accuracy of A3 is due to the use of exceptional visual features provided from ConvNeXt, A3 continues to show poorer balanced accuracy, F1-score, and reliability metrics when compared to DermaFusionNet. These findings demonstrate the value of supplementary diagnostic information such as age, sex, and localization of the lesion that cannot be accurately determined from the image alone. The comparative performance of DermaFusionNet and its ablation variants across multiple evaluation metrics is illustrated in Figure *4*. The ablation study convincingly shows that both gated cross-modal fusion and auxiliary modal regularization are necessary components for achieving clinically relevant performance in the classification of skin diseases.

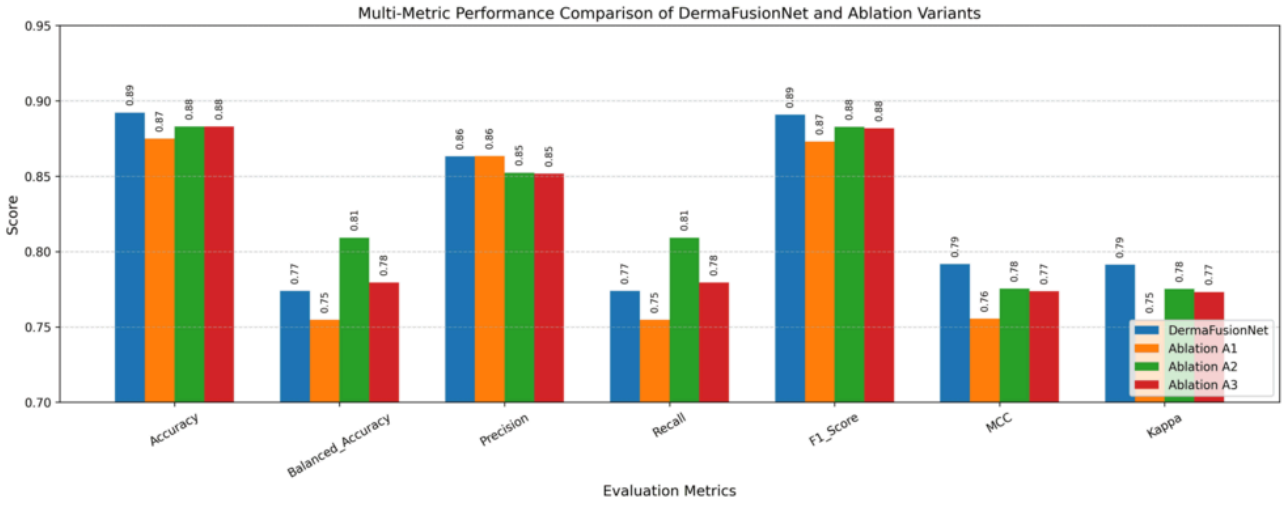

FIGURE 4: Overall performance comparison of DermaFusionNet and its ablation variants across multiple evaluation metrics, including accuracy, balanced accuracy, precision, recall, F1-score, Matthews correlation coefficient (MCC), and Cohen's kappa on the HAM10000 test set. The complete DermaFusionNet model consistently achieves superior performance, demonstrating the effectiveness of gated cross-modal fusion and auxiliary modal regularization

Class-Wise Performance Comparison

Table *7* contains a detailed breakdown of the performance of each class for both the proposed DermaFusionNet and its ablation variants compared to one another. Overall, the complete model shows superior performance across both the majority and minority lesion classes due to the use of an effective gating method to combine the mutually supportive features from both modalities.

Model	Class	Precision	Recall	F1-score	Support
Full Model	akiec	0.867	0.531	0.658	49
	bcc	0.868	0.857	0.863	77
	bkl	0.766	0.794	0.78	165
	df	0.923	0.706	0.8	17
	mel	0.728	0.754	0.741	167
	nv	0.942	0.956	0.949	1,006
	vasc	0.947	0.818	0.878	22
Ablation A1	akiec	0.833	0.612	0.706	49
	bcc	0.846	0.857	0.852	77
	bkl	0.776	0.691	0.731	165
	df	1	0.706	0.828	17
	mel	0.665	0.689	0.676	167
	nv	0.924	0.955	0.939	1,006
	vasc	1	0.773	0.872	22
Ablation A2	akiec	0.814	0.714	0.761	49
	bcc	0.917	0.857	0.886	77
	bkl	0.727	0.806	0.764	165
	df	0.867	0.765	0.813	17
	mel	0.704	0.671	0.687	167
	nv	0.938	0.942	0.94	1,006
	vasc	1	0.909	0.952	22
Ablation A3	akiec	0.909	0.612	0.732	49
	bcc	0.873	0.805	0.838	77
	bkl	0.743	0.824	0.782	165

	df	0.929	0.765	0.839	17
	mel	0.717	0.683	0.699	167
	nv	0.933	0.948	0.941	1,006
	vasc	0.857	0.818	0.837	22

TABLE 7: Class-wise precision, recall, F1-score, and support comparison of DermaFusionNet and its ablation variants on the HAM10000 test set

DermaFusionNet achieves great results for very highly represented classes, with an overall high precision (0.942) and excellent recall rate (0.956), resulting in the overall best F1-score (0.949) across all variants, thus demonstrating superiority in discriminative ability while also proving to be very robust under conditions of class imbalance.

The removal of auxiliary modal regulation in A1 increased the classification precision for some classes (df and vasc) to a fair degree, but also diminished by an equal degree the stability of its recall performance for mel and bkl. This occurred due to a high/low ratio of the dominant cue modulation features. The results of this ablation indicate that A2 had improved recall for some classes when ungated features were concatenated; however, the performance variability reinforces the need for an integrated gating approach in multimodal classifier learning and integration.

Comparable results were achieved by the image-only baseline (Ablation A3) on visually distinctive lesions such as df and bkl. However, consistently lower performance was observed when predicting clinically ambiguous classes such as mel and akiec. This confirms that dermoscopic imaging alone does not provide sufficient evidence for accurate diagnostic performance, and that structured clinical metadata provides essential additional context for diagnostic consideration.

The results of the class-wise assessment of the product indicate that DermaFusionNet is a more reliable and consistent performer across a wide variety of lesions than all of the other products being tested, indicating that it is a valid product to use for supporting decision-making in the routine practice of dermatology.

Confusion Matrix Analysis

As shown in Figure *5*, the full DermaFusionNet model achieves superior class-level discrimination compared to its ablation variants (Figures *6-8*).

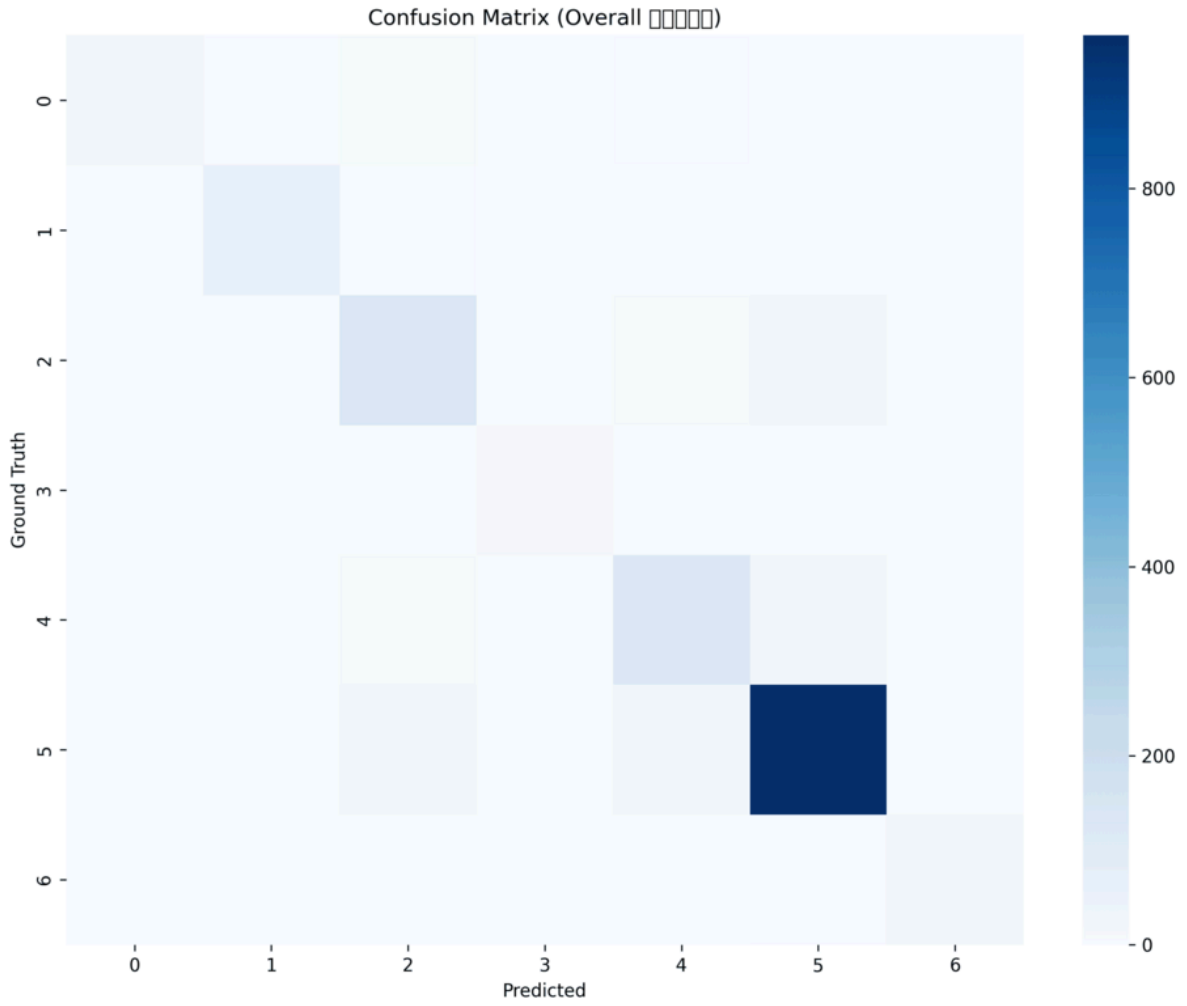

FIGURE 5: Confusion matrix of the proposed DermaFusionNet on the HAM10000 test set, demonstrating strong diagonal dominance and improved class separability across all lesion categories

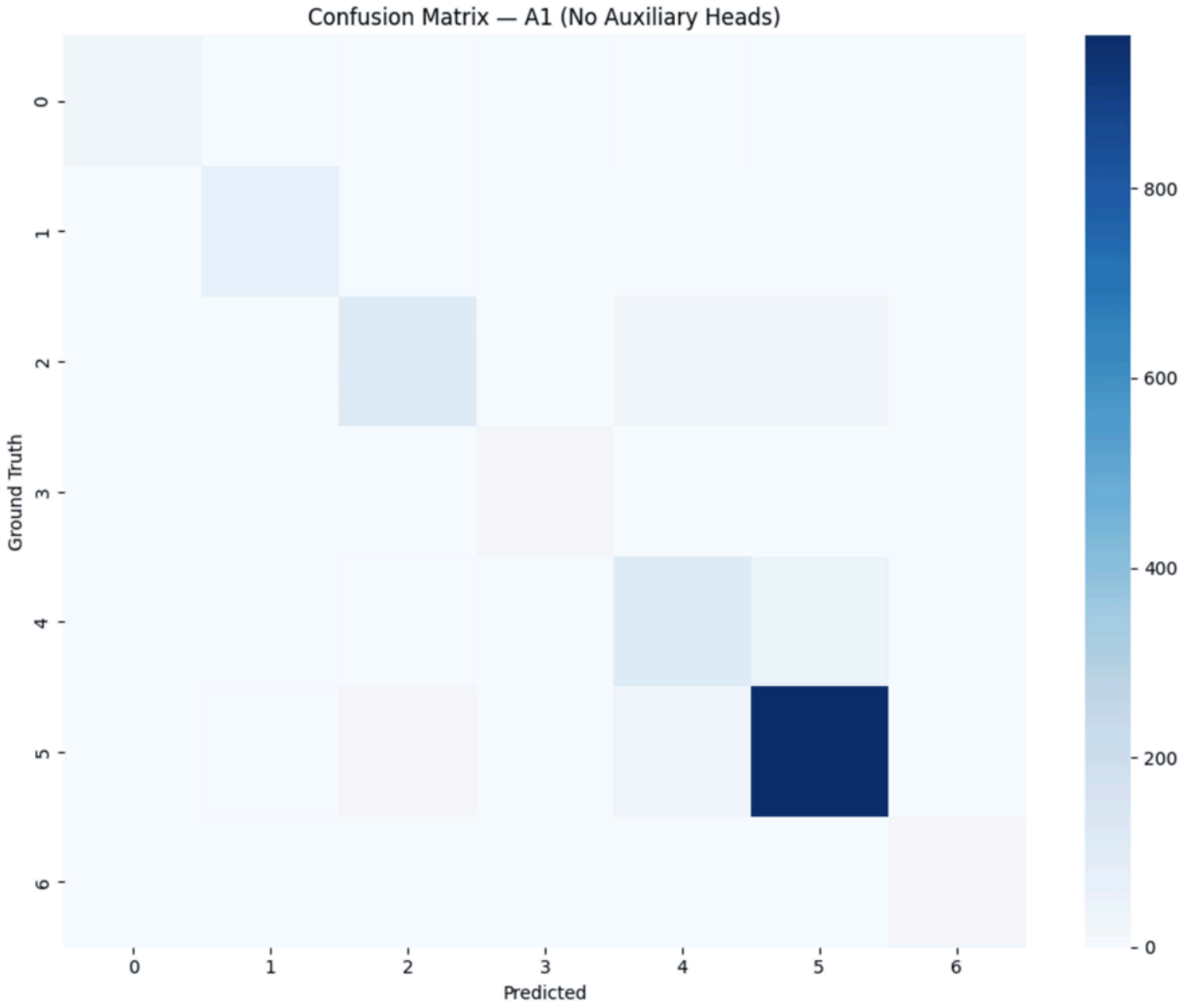

FIGURE 6: Confusion matrix for Ablation A1, where auxiliary modal regularization is removed, resulting in increased misclassification among visually similar lesion classes

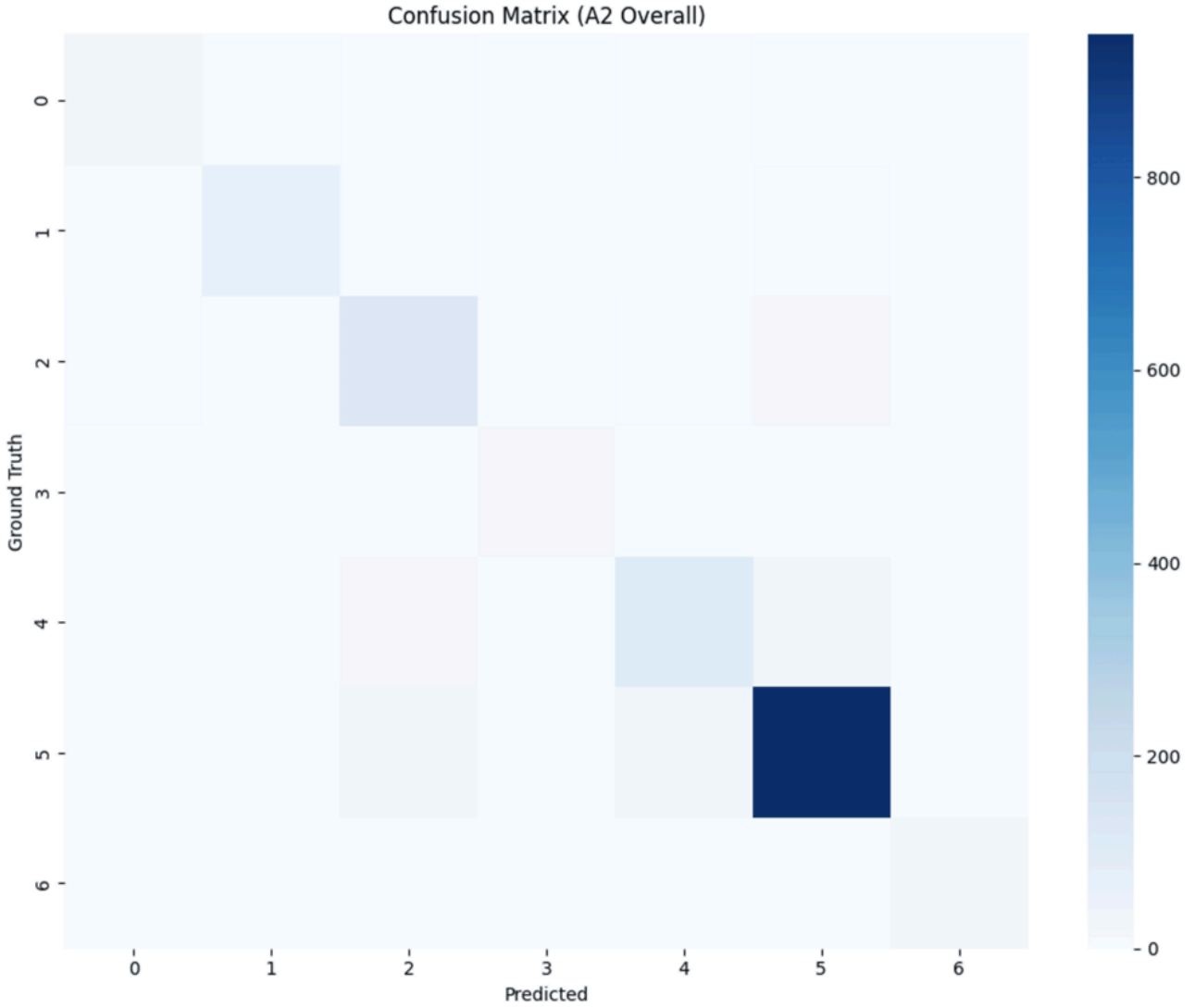

FIGURE 7: Confusion matrix for Ablation A2 employing simple feature concatenation without gating, highlighting uncontrolled feature dominance and reduced discriminative precision

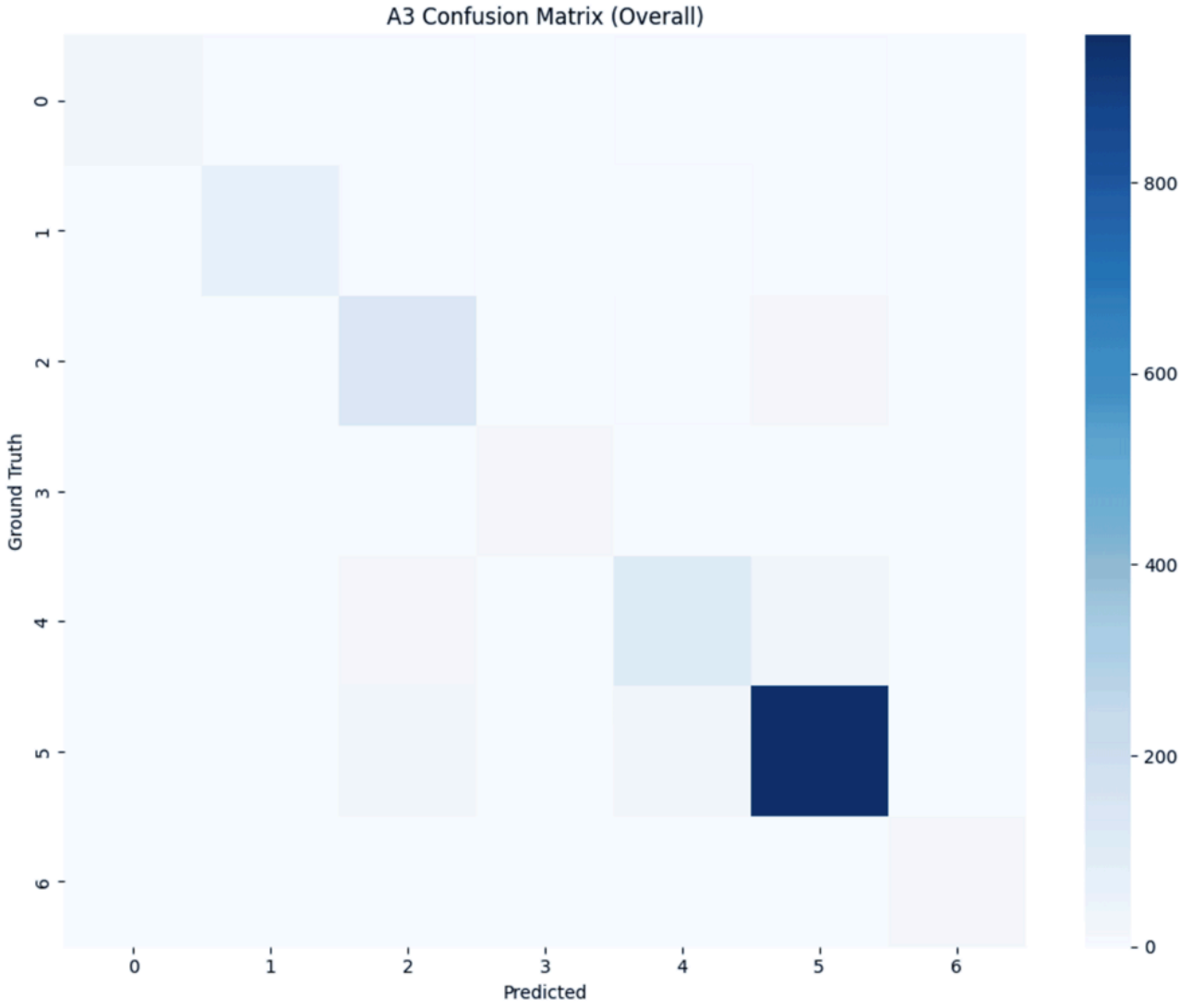

FIGURE 8: Confusion matrix of the image-only model (Ablation A3), illustrating increased confusion for clinically ambiguous lesion categories due to the absence of metadata

Comparison With State-of-the-Art Methods

To validate that DermaFusionNet framework is effective as an option for classifying skin diseases, we compared the DermaFusionNet framework to some recently published state-of-the-art skin disease classification methods using literature published on the subject area. These methods include both deep learning-based approaches and hybrid machine learning models that utilize handcrafted features, segmentation, or ensemble methods to classify skin diseases. This comparison was done using commonly used metrics to measure performance that were reported for all the models presented in Table *8*.

Method	Accuracy	Precision	Recall	F1-score
DermaFusionNet (Proposed)	0.8922	0.8927	0.8922	0.8908
Liu et al. [35]	0.7760	0.6850	0.6420	0.6540
ConvNeXt-Small [37]	0.8400	0.7100	0.8400	0.7200
ResNet50V2 [47]	0.3863	0.4872	0.3745	0.4057
SVM+Texture Features [48]	0.8530	0.8610	0.8350	0.8480
U-Net+SVM [49]	0.8520	0.8300	0.8500	0.8400

TABLE 8: Comparison of DermaFusionNet with state-of-the-art skin disease classification methods on dermoscopic datasets

ResNet, Residual Networks; SVM, Support Vector Machine

Figure *9* provides a visual comparison of the evaluated methods across all performance metrics, highlighting the consistent superiority of DermaFusionNet.

The results indicate that DermaFusionNet is superior to all the other systems evaluated with respect to accuracy, precision, recall, and F1-score. DermaFusionNet has a significant advantage over models that are based solely on images; ResNet50V2 [47] is one example, highlighting that using only images to provide dermatology diagnoses can only be so effective. These results suggest that while hybrid pipelines that combine segmentation and classical classification techniques perform competitively, the reliance on multi-stage procedures to process input data results in the need for pre-designed features, which makes them unscalable and less robust.

On the other end of the spectrum, the system developed by DermaFusionNet connects dermoscopic images and structured clinical data through a gated cross-modal fusion approach, allowing for dynamic weighting of the various features and providing better contextualized reasoning for each individual lesion. Additionally, the very high level of recovery provided by the proposed model is critical for clinical applications and lowers the probability of missing malignant or high-risk lesions. Taken together, the state-of-the-art comparisons confirm that using multimodal data in concert with controlled interaction mechanisms will produce a more usable and reliable clinical tool for the automatic classification of skin diseases.

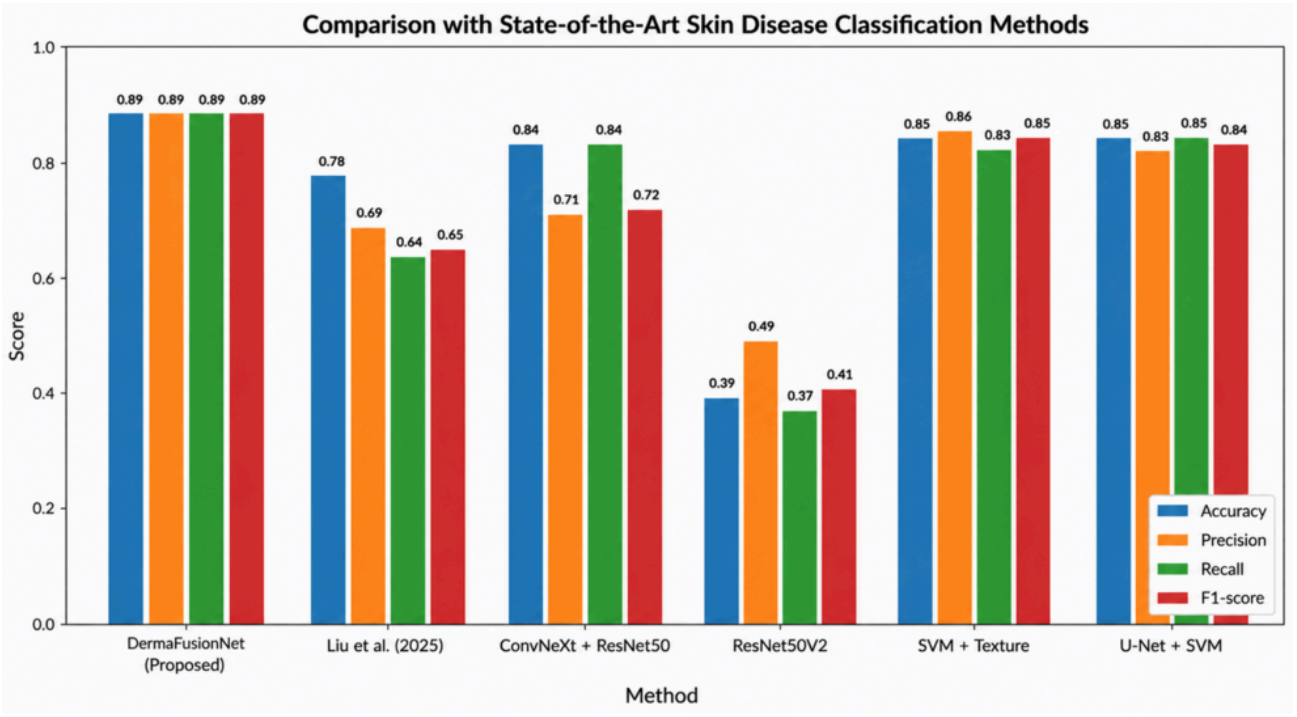

FIGURE 9: Performance comparison of DermaFusionNet with state-of-the-art skin disease classification methods across accuracy, precision, recall, and F1-score

ResNet, Residual Networks; SVM, Support Vector Machine

Explainability and model interpretation

DermaFusionNet performs well quantitatively, but clinical application also requires interpretability and transparency. The general assumption is that deep CNNs are black boxes, preventing clinicians from trusting automated diagnostic systems. To aid in this area, we utilized Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize the spatial areas contributing most to correct predictions. In this part of the study, we present an in-depth explainability analysis, including the setup of Grad-CAM, correct prediction behavior, failure cases, and an overview on the trustworthiness of the model.

Grad-CAM Visualization Setup

In the final convolutional block of the ConvNeXt visual backbone, Grad-CAM was applied to preserve the spatial resolution while preserving high-level semantic representations simultaneously. Grad-CAM produces a localization heatmap by calculating the differentiation of the class score of an input image and its predicted class based on the feature maps. The resulting heatmap indicates where within the feature maps the most influential features of the input image corresponded to the predicted class.

Given feature map A^k and class score y^c, the Grad-CAM weights are computed as:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{7}$$

where Z denotes the total number of spatial locations. The final Grad-CAM map is obtained as:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right) \tag{8}$$

Analysis of Correct Predictions

Figures *10-16* present Grad-CAM visualizations showing images of diseases that our model classifies accurately. Each group contains one example image representing a specific disease category. The images demonstrate that the model consistently focuses on regions of the image that are relevant to diagnosis or diseased areas.

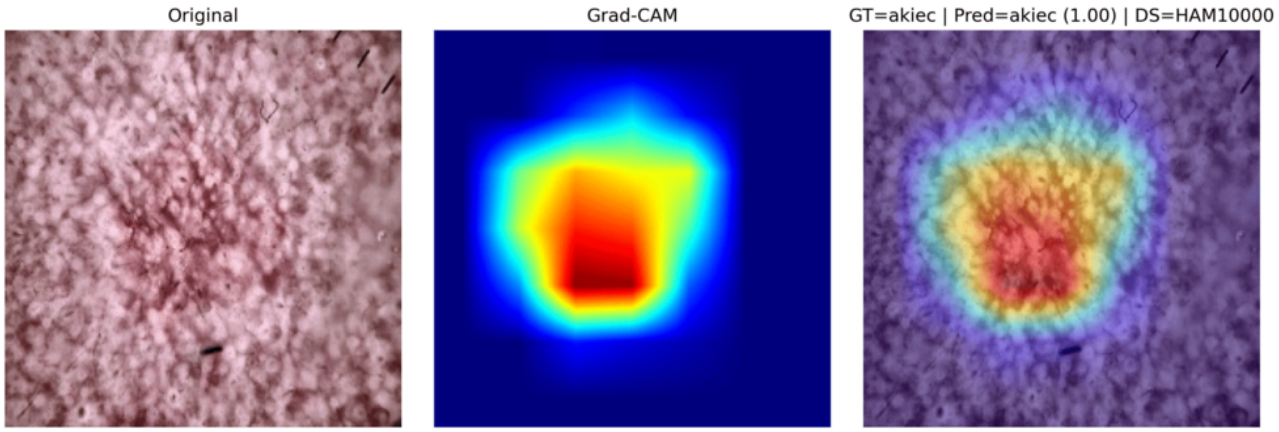

FIGURE 10: Grad-CAM visualization for a correctly classified actinic keratosis (akiec) sample

Grad-CAM, Gradient-weighted Class Activation Mapping

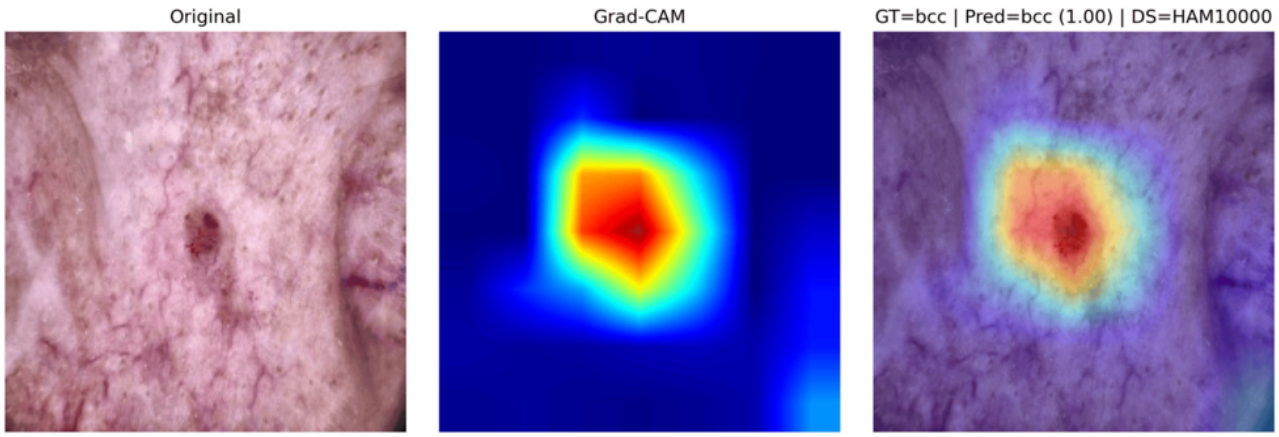

FIGURE 11: Grad-CAM visualization for a correctly classified basal cell carcinoma (bcc) sample

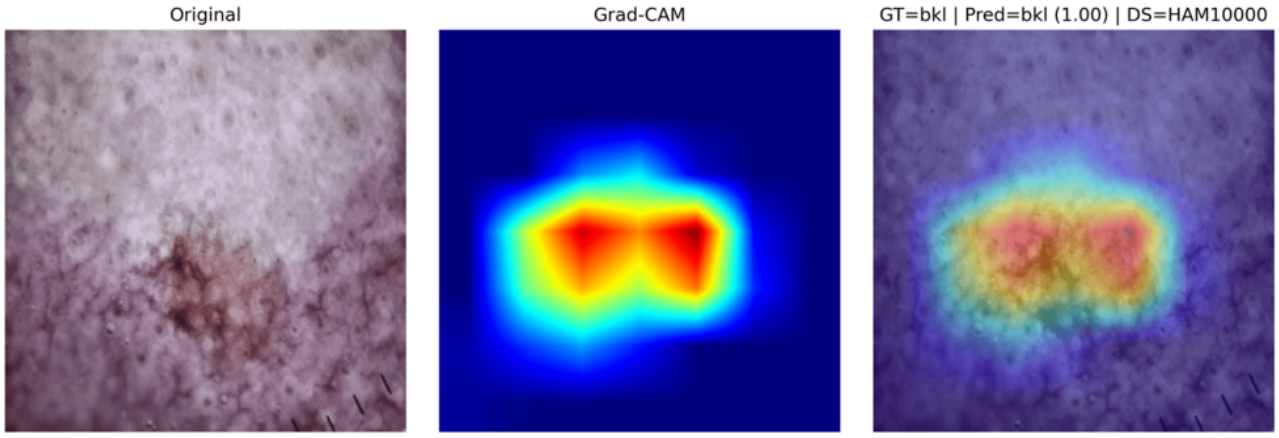

FIGURE 12: Grad-CAM visualization for a correctly classified benign keratosis-like lesion (bkl) sample

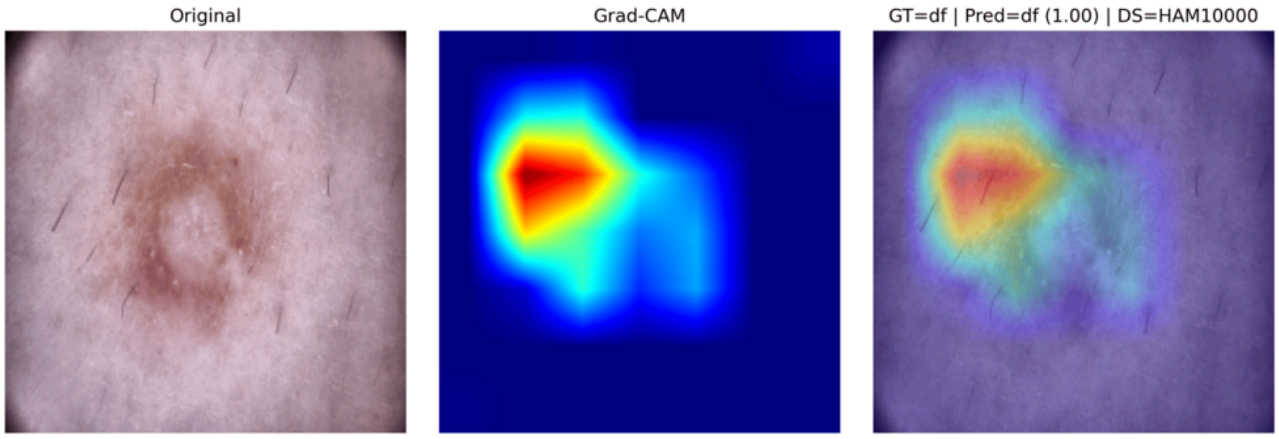

FIGURE 13: Grad-CAM visualization for a correctly classified dermatofibroma (df) sample

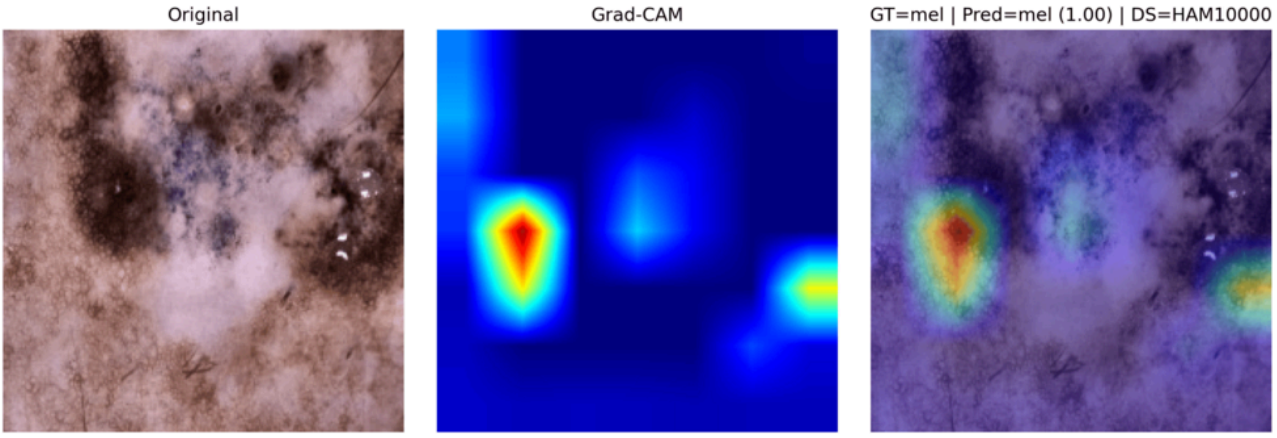

FIGURE 14: Grad-CAM visualization for a correctly classified melanoma (mel) sample

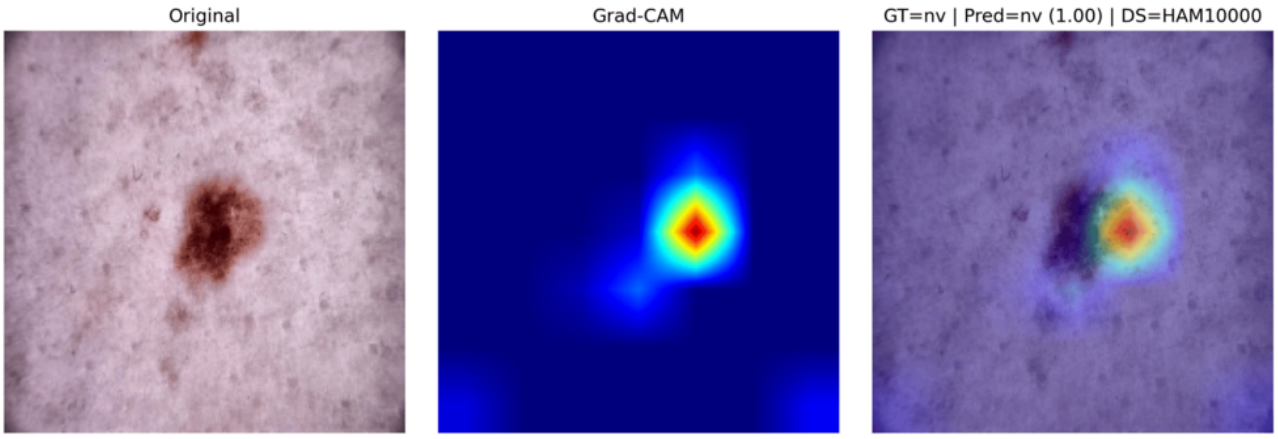

FIGURE 15: Grad-CAM visualization for a correctly classified melanocytic nevus (nv) sample

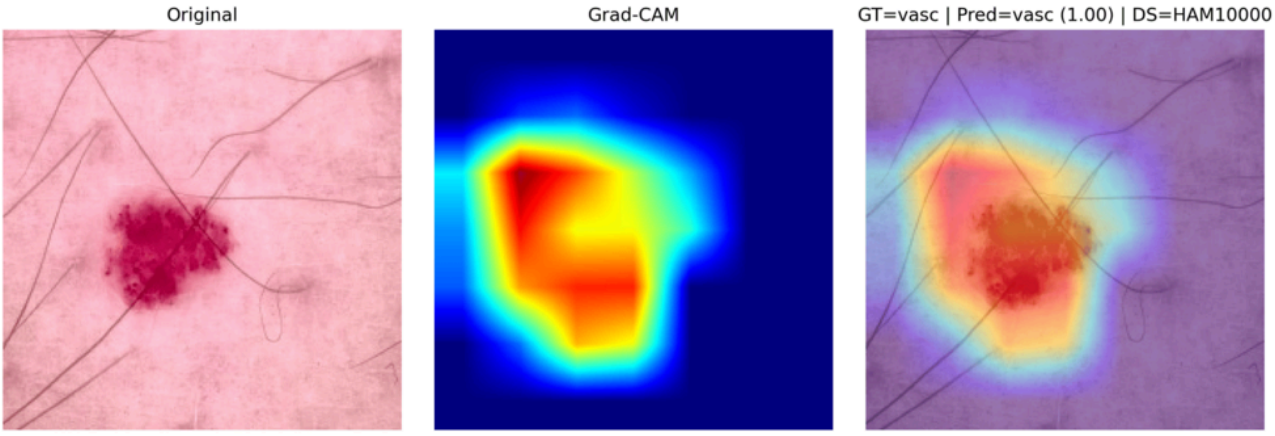

FIGURE 16: Grad-CAM visualization for a correctly classified vascular lesion (vasc) sample

Discussion on Model Trustworthiness

The analyzed Grad-CAM visuals provide evidence that the model is making predictions based on appropriate visual cues in a clinical context. The model makes accurate predictions and displays lesion-based attention patterns. By utilizing both multimodal fusion and clear visual explanations, DermaFusionNet enables improved interpretability and can serve as a reliable component in a clinical decision support system in dermatology.

Discussion

In summary, this research confirms that an effective way of diagnosing skin diseases is to combine dermoscopic images with structured clinical metadata through gated cross-modal fusion. The DermaFusionNet model is superior to both unimodal (image-only) and multimodal baselines in terms of performance and interpretability for automated skin disease classification. This study supports the notion that using multimodal data and controlling the interaction between features from different modalities improves the robustness of dermatological diagnosis.

The quantitative assessments demonstrate DermaFusionNet's excellent balance of both sensitivity and specificity, evidenced by performance metrics (accuracy, weighted F1-score, and MCC). Although the absolute improvement over the image-only baseline in overall accuracy and weighted F1-score appears numerically modest, this should be interpreted in the context of a highly optimized ConvNeXt baseline and the complexity of the HAM10000 benchmark. In clinically sensitive diagnostic systems, even incremental performance gains can be meaningful when they are accompanied by improved balanced accuracy, stronger minority-class discrimination, and enhanced robustness for visually ambiguous lesion categories. Therefore, the practical significance of DermaFusionNet lies not only in aggregate metric improvement but also in its ability to provide more context-aware, reliable, and clinically informed predictions through multimodal reasoning. In comparison to the image-only variant of the model (Ablation A3), the complete model leverages the additional information contained in the metadata and has an even larger impact on enhancing the accuracy of ambiguous lesion types (such as actinic keratosis and benign keratosis-like lesions). These results illustrate the ability of clinical information (e.g., lesion localization, patient demographics, and acquisition source) to help mitigate the visual uncertainty associated with dermoscopic images.

The ablation analysis illustrates the significance of architectural choices. The removal of auxiliary modal regularization (the first ablation) greatly affects both balanced accuracy and class-wise recall, as they decrease significantly. Auxiliary constraints stabilize multimodal feature alignment during training and prevent this drop in performance. When gated fusion is replaced with simple concatenation (the second ablation), even greater drops in performance, especially for

minority classes, occur, indicating that unguided feature merging introduces modality dominance and decreases the discriminative capability of the overall model. Collectively, these results support the need for both forms of regularization (gated and auxiliary) in order to achieve optimal performance.

The class-wise analysis shows that DermaFusionNet is superior in classifying the most common lesions, such as melanocytic nevus and melanoma, compared to less frequent lesion types, such as dermatofibromas and vascular lesions. Most remaining misclassifications occur between lesions with similar clinical appearances. This finding is consistent with the current understanding of the difficulties dermatologists face in diagnosing skin lesions. These misclassifications do not occur uniformly across all classes; rather, they are primarily clustered in classes that exhibit subtle visual differences or are separated by very little inter-class contrast.

Explainability analysis via Grad-CAM provides additional insight into model behavior and reliability in correct predictions. Through the activation maps, it is evident that the lesion is consistently highlighted in line with current clinical standards. The model's attention remains on lesions in areas of failure cases rather than background artifacts, indicating that such misclassifications would result from the intrinsic visual ambiguity of the model rather than spurious correlations. The model's decision-making process therefore increases confidence in its potential as a clinical decision support system, compared to a black-box classifier.

The superior performance of multimodal fusion methods relative to current techniques, as evidenced in the comparison with state-of-the-art methods, demonstrates that these techniques will provide better outcomes than either visual or hybrid methods developed with human ingenuity. Classical methods that integrate machine learning with segmentation provide adequate results but will require multiple processing stages and a high degree of feature engineering. On the other hand, DermaFusionNet provides a full end-to-end learning environment with the ability to simultaneously optimize both visual information and clinical data for better generalization and better ability to interpret results.

There are some limitations associated with the strengths of the model that should be addressed. Because the training and evaluation of the model were performed using one gold standard source of data, the potential for dataset-specific bias exists regarding how images were obtained and annotated. Furthermore, while integrating clinical metadata improved overall accuracy, the availability and quality of this information in real clinical environments can be inconsistent and incomplete, making the need to address these challenges imperative if the model is going to be reliably applied in a variety of clinical environments. Although HAM10000 provides a widely accepted benchmark for dermoscopic skin lesion analysis, evaluation on a single dataset may limit the generalizability of the proposed framework across broader clinical environments, acquisition settings, and patient populations. While the use of stratified held-out testing supports internal robustness, future work should include external validation on independent datasets such as ISIC, PH2, or PAD-UFES to further assess cross-domain stability, demographic generalization, and real-world clinical applicability.

The results of this study demonstrate the critical nature of using principled multimodal design, controlled feature fusion, and explainability in order to conduct medical image analysis. The DermaFusionNet is an important advancement to create an automated process to classify skin disease that is reliable, transparent, and clinically applicable.

Conclusions

This paper presented DermaFusionNet, a multimodal deep learning framework for skin disease classification that integrates dermoscopic images with structured clinical metadata through gated cross-modal fusion. By jointly learning visual and contextual representations, the proposed model effectively addresses the inherent ambiguity of dermoscopic imagery and improves diagnostic reliability. Beyond aggregate performance gains, the proposed gated fusion strategy enhances contextual adaptability and multimodal robustness, supporting more reliable deployment in clinically heterogeneous environments. Extensive experiments on the HAM10000 dataset demonstrate that DermaFusionNet consistently outperforms image-only and ungated multimodal baselines, achieving strong accuracy, precision, recall, and F1-score while maintaining robust performance across both common and underrepresented lesion categories. Furthermore, explainability analysis using Grad-CAM confirms that the model's predictions are driven by clinically meaningful lesion regions, reinforcing its suitability as a transparent and trustworthy decision-support system.

Future work will be dedicated to developing the Derma-FusionNet architecture for the evaluation of multiple datasets and the cross-domain generalization of multiple datasets so that it performs well under various acquisition conditions and patient demographic groups. Implementing more sophisticated methods of balancing classes, such as using cost-sensitive loss functions or adaptive sampling methods, will assist with class imbalance issues as well as ensuring that rare diseases have adequate representation within the class distribution of the dataset. Other avenues for further investigation include developing or employing transformer-like fusion mechanisms, integrating higher-resolution dermoscopy images, and using the collection of longitudinal data associated with patients in order to analyze disease progressions and facilitate their implementation in the real-world clinical setting.

Additional Information

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

All data generated or analyzed during this study are included in this published article and/or its appendices.

References

1. Grant SR, Andrew TW, Alvarez EV, Huss WJ, Paragh G: Diagnostic and prognostic deep learning applications for histological assessment of cutaneous melanoma. Cancers. 2022, 14:6231. 10.3390/cancers14246231
2. Bassel A, Abdulkareem AB, Alyasseri ZAA, Sani NS, Mohammed HJ: Automatic malignant and benign skin cancer classification using a hybrid deep learning approach. Diagnostics. 2022, 12:2472. 10.3390/diagnostics12102472
3. Kaur R, GholamHosseini H, Sinha R, Lindén M: Melanoma classification using a novel deep convolutional neural network with dermoscopic images. Sensors. 2022, 22:1134. 10.3390/s22031134
4. Naeem A, Anees T, Fiza M, Naqvi RA, Lee S-W: SCDNet: a deep learning-based framework for the multiclassification of skin cancer using dermoscopy images. Sensors. 2022, 22:5652. 10.3390/s22155652
5. Tahir M, Naeem A, Malik H, Tanveer J, Naqvi RA, Lee S-W: DSCC_Net: multi-classification deep learning models for diagnosing of skin cancer using dermoscopic images. Cancers. 2023, 15:2179. 10.3390/cancers15072179
6. Liu L, Liang C, Xue Y, et al.: An intelligent diagnostic model for melasma based on deep learning and multimode image input. Dermatology and Therapy. 2023, 13:569-579. 10.1007/s13555-022-00874-z
7. Fraiwan M, Faouri E: On the automatic detection and classification of skin cancer using deep transfer learning. Sensors. 2022, 22:4963. 10.3390/s22134963
8. Naeem A, Anees T: A multiclassification framework for skin cancer detection by the concatenation of Xception and ResNet101. Journal of Computing & Biomedical Informatics. 2024, 6:205-227.
9. Saleh N, Hassan MA, Salaheldin AM: Skin cancer classification based on an optimized convolutional neural network and multicriteria decision-making. Scientific Reports. 2024, 14:17323. 10.1038/s41598-024-67424-9
10. Alsahafi YS, Kassem MA, Hosny KM: Skin-Net: a novel deep residual network for skin lesions classification using multilevel feature extraction and cross-channel correlation with detection of outlier. Journal of Big Data. 2023, 10:105. 10.1186/s40537-023-00769-6

11. Farag TK, Abou-Zeina HAA, Abdel-Shafy S, Allam AM, Ghazy AA: Progress in diagnostic methods and vaccines for lumpy skin disease virus: a path towards understanding the disease. Veterinary Research Communications. 2025, 49:134. 10.1007/s11259-025-10667-2

12. Sivakumar N, Khan AR, Umar S, et al.: A hybrid brain tumor classification using FL with FedAvg and FedProx for privacy and robustness across heterogeneous data sources. IEEE Access. 2025, 13:57705-57719. 10.1109/access.2025.3549440

13. Bechelli S, Delhommelle J: Machine learning and deep learning algorithms for skin cancer classification from dermoscopic images. Bioengineering. 2022, 9:97. 10.3390/bioengineering9030097

14. Patil R, Bellary S: Machine learning approach in melanoma cancer stage detection. Journal of King Saud University Computer and Information Sciences. 2022, 34:3285-3293. 10.1016/j.jksuci.2020.09.002

15. Marks R: An overview of skin cancers. Cancer. 1995, 75:607-612. 10.1002/1097-0142(19950115)75:2+<607::AID-CNCR2820751402>3.0.CO;2-8

16. Chauhan P, Lunagaria M, Verma DK, Vaghela K, Tejani GG, Sharma SK, Khan AR: PBVit: A patch-based vision transformer for enhanced brain tumor detection. IEEE Access. 2025, 13:13015-13029. 10.1109/access.2024.3521002

17. Hosny KM, Kassem MA: Refined residual deep convolutional network for skin lesion classification. Journal of Digital Imaging. 2022, 35:258-280. 10.1007/s10278-021-00552-0

18. Vora C, Katkar V, Lunagaria M: Gait analysis based on gender detection using pre-trained models and tune parameters. Discover Artificial Intelligence. 2024, 4:19. 10.1007/s44163-024-00115-6

19. Kumar A, Vatsa A: Untangling classification methods for melanoma skin cancer. Frontiers in Big Data. 2022, 5:848614. 10.3389/fdata.2022.848614

20. Lunagaria M, Katkar V, Vaghela K: Covid-19 and pneumonia infection detection from chest X-ray images using U-Net, EfficientNetB1, XGBoost and recursive feature elimination. International Journal of Advanced Computer Science and Applications. 2022, 13:475-483.

21. Naeem A, Anees T: DVFNet: A deep feature fusion-based model for the multiclassification of skin cancer utilizing dermoscopy images. PLoS ONE. 2024, 19:e0297667. 10.1371/journal.pone.0297667

22. Saha U, Ahamed IU, Imran MA, Ahamed IU, Hossain A, Gupta UD: YOLOv8-based deep learning approach for real-time skin lesion classification using the HAM10000 dataset. 2024 IEEE International Conference on E-health Networking, Application & Services (HealthCom), Nara, Japan. 2024, 1-4. 10.1109/HealthCom60970.2024.10880715

23. Groh M, Badri O, Daneshjou R, et al.: Deep learning-aided decision support for diagnosis of skin disease across skin tones. Nature Medicine. 2024, 30:573-583. 10.1038/s41591-023-02728-3

24. Reddy DA, Roy S, Kumar S, Tripathi R: A scheme for effective skin disease detection using optimized region growing segmentation and autoencoder based classification. Procedia Computer Science. 2023, 218:274-282. 10.1016/j.procs.2023.01.009

25. Sadik R, Majumder A, Biswas AA, Ahammad B, Rahman MM: An in-depth analysis of convolutional neural network architectures with transfer learning for skin disease diagnosis. Healthcare Analytics. 2023, 3:100143. 10.1016/j.health.2023.100143

26. Thapar P, Rakhra M, Alsaadi M, Quraishi A, Deka A, Ramesh JVN: A hybrid Grasshopper optimization algorithm for skin lesion segmentation and melanoma classification using deep learning. Healthcare Analytics. 2024, 5:100326. 10.1016/j.health.2024.100326

27. Naeem A, Anees T, Khalil M, Zahra K, Naqvi RA, Lee S-W: SNC_Net: skin cancer detection by integrating handcrafted and deep learning-based features using dermoscopy images. Mathematics. 2024, 12:1030. 10.3390/math12071030

28. Lunagariya M, Katkar V: Light weight approach for COVID-19, pneumonia detection from X-ay images using deep feature extraction and XGBoost. 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India. 2022, 1-5. 10.1109/ICCSEA54677.2022.9936194

29. Naeem A, Khan AH, Ayubi SUD, Malik H: Predicting the metastasis ability of prostate cancer using machine learning classifiers. Journal of Computing & Biomedical Informatics. 2023, 4:1-7.

30. Adebiyi A, Abdalnabi N, Smith EH, Hirner J, Simoes EJ, Becevic M, Rao P: Accurate skin lesion classification using multimodal learning on the HAM10000 dataset [PREPRINT]. medRxiv. 2024, 10.1101/2024.05.30.24308213

31. Riaz S, Naeem A, Malik H, Naqvi RA, Loh W-K: Federated and transfer learning methods for the classification of melanoma and nonmelanoma skin cancers: a prospective study. Sensors. 2023, 23:8457. 10.3390/s23208457

32. Chiu M-C, Wang Y, Kuo Y-J, Chen P-Y: DDI-CoCo: A dataset for understanding the effect of color contrast in machine-assisted skin disease detection. ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea. 2024, 6905-6909. 10.1109/ICASSP48485.2024.10448011

33. AlSadhan NA, Alamri SA, Ben Ismail MM, Bchir O: Skin cancer recognition using unified deep convolutional neural networks. Cancers. 2024, 16:1246. 10.3390/cancers16071246

34. Thanga Purni JS, Vedhapriyavadhana R: EOSA-Net: A deep learning framework for enhanced multi-class skin cancer classification using optimized convolutional neural networks. Journal of King Saud University Computer and Information Sciences. 2024, 36:102007. 10.1016/j.jksuci.2024.102007

35. Liu H, Dou Y, Wang K, Zou Y, Sen G, Liu X, Li H: A skin disease classification model based on multi scale combined efficient channel attention module. Scientific Reports. 2025, 15:6116. 10.1038/s41598-025-90418-0

36. Behara K, Bhero E, Agee JT: Grid-based structural and dimensional skin cancer classification with self-featured optimized explainable deep convolutional neural networks. International Journal of Molecular Sciences. 2024, 25:1546. 10.3390/ijms25031546

37. Gairola AK, Kumar V, Sahoo AK, Diwakar M, Singh P, Garg D: Multi-feature fusion deep network for skin disease diagnosis. Multimedia Tools and Applications. 2025, 84:419-444. 10.1007/s11042-024-18958-7

38. Grossarth S, Mosley D, Madden C, Ike J, Smith I, Huo Y, Wheless L: Recent advances in melanoma diagnosis and prognosis using machine learning methods. Current Oncology Reports. 2023, 25:635-645. 10.1007/s11912-023-01407-3

39. Słowińska M, Czarnecka I, Czarnecki R, et al.: Characteristics of patients with melanoma with non-melanoma skin cancer comorbidity: Practical implications based on a retrospective study. Oncology Letters. 2025, 29:1-14. 10.3892/ol.2025.14960

40. Karp P, Karp K, Kądziela M, Zajdel R, Żebrowska A: The importance of early detection and prevention of atypical skin lesions and other melanoma risk factors in a younger population. Cancers. 2024, 16:4264. 10.3390/cancers16244264

41. Naseri H, Safaei AA: Diagnosis and prognosis of melanoma from dermoscopy images using machine learning and deep learning: a systematic literature review. BMC Cancer. 2025, 25:75. 10.1186/s12885-024-13423-y

42. Sánchez Moreno EC, Machado Sulbaran AC, Leal LR, Ortiz García YM, Olivas Román LR, Riera Leal A: Non-melanoma skin cancer: Dermatoscopic diagnostic clues in mexican individuals based on fitzpatrick skin phototypes. Journal of Clinical Medicine. 2025, 14:2966. 10.3390/jcm14092966

43. Unnisa Z, Tariq A, Sarwar N, Din I, Serhani MA, Trabelsi Z: Impact of fine-tuning parameters of convolutional neural network for skin cancer detection. Scientific Reports. 2025, 15:14779. 10.1038/s41598-025-99529-0

44. Gautam V, Trivedi NK, Anand A, Tiwari R, Zaguia A, Koundal D, Jain S: Early skin disease identification using deep neural network. Computer Systems Science and Engineering. 2023, 44:2259-2275. 10.32604/csse.2023.026358

45. Lyakhov PA, Lyakhova UA, Kalita DI: Multimodal analysis of unbalanced dermatological data for skin cancer recognition. IEEE Access. 2023, 11:131487-131507. 10.1109/access.2023.3336289

46. Tschandl P, Rosendahl C, Kittler H: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data. 2018, 5:180161. 10.1038/sdata.2018.161

47. Balasundaram A, Shaik A, Rohan Alroy B, Singh A, Shivaprakash SJ: Genetic algorithm optimized stacking approach to skin disease detection. IEEE Access. 2024, 12:88950-88962. 10.1109/ACCESS.2024.3412791

48. Das K, Cockerell CJ, Patil A, Pietkiewicz P, Giulini M, Grabbe S, Goldust M: Machine learning and its application in skin cancer. International Journal of Environmental Research and Public Health. 2021, 18:13409. 10.3390/ijerph182413409

49. Seeja RD, Suresh A: Deep learning based skin lesion segmentation and classification of melanoma using support vector machine (SVM). Asian Pacific Journal of Cancer Prevention. 2019, 20:1555-1561. 10.31557/apjcp.2019.20.5.1555