Comparing Image-Clinical Data Fusion Approaches for Skin Lesion Classification using Deep Learning

Adriteyo Das [1,*], Vedant Agarwal [2] and Nisha P Shetty [1,†]

[1] *Department of Information and Communication Technology*
[2] *Department of Humanities and Management*
Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India

Correspondence*:
Nisha P Shetty
nisha.pshetty@manipal.edu

ABSTRACT

Skin lesion classification is a critical task in dermatology, with significant implications for early diagnosis and treatment. In this study, we propose a novel multimodal data fusion approach that integrates dermatoscopic images and clinical metadata to improve classification accuracy. We systematically evaluate various fusion techniques, including simple concatenation, weighted concatenation, and advanced methods such as self-attention and cross-attention, using the widely recognized HAM10000 dataset. Our results demonstrate that combining clinical features with image data significantly enhances classification performance, with cross-attention fusion achieving the highest accuracy because it effectively captures inter-modal dependencies and contextual relationships between different data modalities. Furthermore, we employ Grad-CAM to improve model interpretability, providing insights into the relevance of clinical features in decision-making. Despite these advancements, challenges such as class imbalance and the computational complexity of advanced fusion methods persist. We also discuss future directions to optimize model efficiency and interpretability for clinical use in resource-constrained environments.

Keywords: skin lesion classification, multimodal fusion, dermatoscopic images, clinical metadata, cross-attention, HAM10000, interpretability, deep learning

1 INTRODUCTION

Skin Cancer is the 5th most prevalent type of cancer. It is one of the most devastating forms and is expected to soon surpass heart disease as the main cause of mortality among humans. [1] Between 1990 and 2017, the prevalence of individuals with Malignant skin melanoma, Squamous cell carcinoma, and Basal cell carcinoma has increased by 215.7%, 196.8%, and 90.9% respectively.[2]. The most common types of skin cancers are Melanoma, Basal Cell Carcinoma (BCC), and Squamous Cell Carcinoma (SCC), along with precancerous conditions like Actinic Keratosis (AK)[3]. Being a leading global health concern, skin cancer underscores the critical need for early diagnosis to improve patient outcomes and reduce mortality. Early detection enables less invasive treatments and lowers healthcare costs by identifying cancers at treatable stages. [4] Current diagnostic methods, such as skin self-examination (SSE) and clinical skin examination (CSE), rely heavily on visual inspection and tools such as the ABCDE rule (Asymmetry, Border, Color, Diameter, Evolution) rule. Although advanced techniques such as dermoscopy and total body photographygy (TBP) improve accuracy, traditional methods remain subjective, time-consuming, and inaccessible in underserved areas [5] [6] [7].

Computerized systems, including Computer-aided Diagnosis (CAD) software and image processing algorithms, offer reproducible, objective, and quick evaluation of skin lesions, less dependent on subjective human judgment. Such systems can process large volumes of data with efficiency, allowing early detection of subtle patterns that would be easily overlooked by conventional techniques (Han et al., 2018). Automation also enhances accessibility by being incorporated into telemedicine platforms, extending diagnostic abilities to distant and under-served communities. By minimizing the necessity for invasive biopsies and follow-up visits, automation decreases healthcare expenditure without compromising diagnostic efficiency and patient outcome. [8]

Machine learning (ML) and deep learning (DL) have made tremendous progress over the last few years, fueled by more computational power and large datasets. These technologies have shown remarkable success in a range of medical classification problems, including eye movement-based disease prediction with Decision Trees and Random Forests, automated skin disease classification with KNN and SVM with 98.22% accuracy, and brain tumor classification with Convolutional Neural Networks (CNNs) such as VGG16 and VGG19 with accuracies ranging from 92.5% to 97.8% [9].

Haenssle et al. [10] compared the diagnostic accuracy of a bespoke Convlutational Neural Network (CNN) on the basis of Google's Inception v4 architecture, which was trained using data from cooperating dermatologists and the International Skin Imaging Collaboration (ISIC) dermoscopic archive, with a large international cohort of 58 dermatologists. The findings showed that the CNN performed better than most dermatologists with a mean AUC-ROC of 0.86 against 0.79 (p ¡ 0.01). This work brings into perspective the promise for dermatologists to incorporate CNNs into their workflow, leading to increased diagnostic precision and eventually enhanced patient outcomes in terms of improved prognosis, treatment, and quality of life.

Multimodal data represents information derived from diverse sources and formats, including images, text, audio, and physiological signals. This integrated approach mirrors human cognitive processes, where multiple sensory modalities contribute to perception and interpretation. In medical contexts, multimodal data combines Medical images, Patient records (demographics, medical history, lab results), Physiological signals, and Patient-reported outcomes.

Recent studies have demonstrated the efficacy of multimodal approaches in medical diagnosis. For example, Zanetti *et al.* (2024) achieved 98.27% accuracy using a Multi-feature Kernel Supervised within-class-similar Discriminative Dictionary Learning (MKSCDDL) algorithm for Alzheimer's Disease classification [11], Jiang *et al.* (2022) employed multimodal ultrasound data with a CNN, achieving 98.22% accuracy in early breast cancer detection [12], and Kumar *et al.* (2022) utilized audio and X-ray imaging with a CNN and Deep Uniform Net, obtaining 98.67% accuracy in COVID-19 classification [13].

In this study, we analyze fusion techniques for Skin cancer classification, leveraging multimodal skin images and clinical data. Our research explores how fusion methods can enhance skin lesion classification performance compared to single-modality approaches. We investigate various fusion strategies to integrate diverse data sources and improve model accuracy.

To enhance the interpretability of our multimodal systems, we apply the grad-CAM approach for deep learning explainability and feature relevance assessment. Our contributions aim to advance skin lesion classification by presenting robust fusion strategies that offer high accuracy and clinical interpretability.

2 RELATED WORK

The automatic detection and classification of skin lesions, particularly for skin cancer diagnosis, has been a critical area of research in Applied AI. Recent studies have explored diverse methodologies, ranging from texture-based feature extraction to multimodal deep learning, aiming to enhance the accuracy, efficiency, and explainability of automated diagnosis systems.

Arshad et al. [14] developed a deep learning-based diagnostic system achieving **95% accuracy** on the **HAM100000 dataset**. Using **ResNet-50 and ResNet-101** with **ensemble subspace discriminative classifiers**, they combined transfer learning, feature extraction, and a modified feature fusion approach. However, their reliance on dermoscopic images and computationally expensive fusion methods limits its real-world application and interpretability.

Sevli et al. [15] proposed a **CNN-based model** for classifying pigmented skin lesions using the **HAM10000 dataset** The model employs preprocessing steps such as **image resizing**, contrast enhancement, and sharpening filters. The CNN architecture comprises three main blocks with convolutional, pooling, dropout, and fully connected layers, and a **Softmax activation function** for multi-class classification. The model achieved **91.51% accuracy** on the test set and **90.28% accuracy** during evaluation with seven expert dermatologists, correcting **11.14% of their misdiagnoses**. It performed best on **Melanocytic Nevi** but struggled with **Dermatofibromas** and often confused **melanoma with Melanocytic Nevi**. However, the study is **not multimodal**, relying solely on image data. The limitations include reduced performance for cases with **hair artifacts** or **dark skin tones**, a **small sample size** for some classes, and the lack of integration of additional features like **color or shape**. Increasing expert evaluations and expanding test datasets could further improve the model's generalizability.

Wang et al. [16] proposed a **two-stream network** leveraging **DenseNet-121** and an improved **VGG-16** with residual structures. This model incorporated a **multireceptive field module** and **Generalized Mean Pooling (GeM)**, achieving **91.24% accuracy** on the **HAM10000 dataset**. Despite its effectiveness, the model's sole reliance on dermoscopic images and its non-lightweight architecture could hinder its deployment in resource-constrained settings.

Adebiyi et al. [17] explored the **ALBEF (Align Before Fuse)** framework for **multimodal skin lesion classification**, integrating images with metadata (age, sex, lesion location). The joint text-image encoder, combining a **Vision Transformer** and **BERT**, achieved a **94.11% accuracy** and **AUC-ROC of 0.9426**, surpassing image-only models. However, the study's reliance on only three metadata variables suggests potential for further improvement with additional clinical features.

Srivastava et al. [18] introduced a texture-based feature extraction technique, **Median-based Quadrilateral Local Ternary Quantized Pattern (M-QuadLTQP)**, achieving **96% accuracy** on the **HAM10000 and ISIC UDA datasets**. By combining median-based noise reduction with a modified CNN, this method showed superior performance over traditional texture analysis techniques. However, its **computational intensity** and **lack of multimodality** limit its scalability and clinical applicability.

Datta et al.[19] investigated the use of a **Soft-Attention mechanism** in deep neural networks for skin lesion classification using the **HAM10000** and **ISIC 2017 datasets**. The Soft-Attention module, integrated into architectures like **Inception ResNet v2**, **ResNet**, and **DenseNet**, generates attention maps via **3D convolution** and **softmax normalization**, enhancing focus on salient features. Their best-performing model, **IRv2+SA**, achieved **93.7% precision** and improved sensitivity by **3.8%**. Despite its transparency claims, the approach is **not multimodal**, lacks generalizability testing across diverse datasets, and has lower specificity, which may increase false positives. The study acknowledges that the generalization performance of FixCaps has not been adequately studied, which will be a focus of future work.

Lan et al. [20] introduced **FixCaps**, an improved capsule network for **dermoscopic image classification**. FixCaps uses a **large-kernel convolution** (31x31) at the bottom layer, enabling a larger receptive field, and incorporates a **Convolutional Block Attention Module (CBAM)** to retain spatial information. The capsule layer includes both **primary and digit capsules**, with group convolution (GP) to avoid underfitting. The study also proposes **FixCaps-DS**, which combines FixCaps with deep-wise separable convolution for mobile deployment. Tested on the **HAM10000 dataset**, FixCaps achieved **96.49% accuracy**, outperforming IRv2-SA. FixCaps-DS showed **96.13% accuracy** with significantly fewer parameters. The study found larger kernels improve feature learning, but the model underperformed on the VASC lesion type. The study acknowledges that the generalization performance of FixCaps has not been adequately studied, which will be a focus of future work. **FixCaps** is not multimodal, as it uses only dermoscopic images for classification without patient metadata.

Gessert et al. [21] achieved first place in the ISIC 2019 Skin Lesion Classification Challenge by using an **ensemble of deep learning models**, including EfficientNets, SENet, and ResNeXt WSL, selected via a search strategy. The approach involved multiple **input resolutions**, **cropping strategies**, and a **loss balancing method** to handle class imbalance. A **patient meta data** branch (age, sex, anatomical site) was incorporated via a dense neural network, improving performance by 1–2% for smaller models. The architecture consisted of EfficientNets pretrained on ImageNet, SENet154, and ResNext variants, along with extensive **data augmentation** (brightness, contrast, flipping, rotation, scaling, shear, CutOut). For Task 2, meta data was processed separately using a two-layer neural network with batch normalization, ReLU, and dropout. **Ensembling** of the models led to substantial performance improvements, with optimal results achieved from nine out of sixteen configurations. However, the performance on the official test set was notably lower than cross-validation, especially for the **unknown class**. **Multimodal aspects** of the study combined **image data** and **meta data** for classification. The limitations of this study include **overfitting** to missing meta data in the unknown class and reduced performance when incorporating this class. The study also found that **mixture of input resolutions** was beneficial, but the model's performance was significantly lower on the unknown class across several metrics.

Ou et al. [22] introduced a **multimodal model** utilizing **smartphone images** and metadata (e.g., age, sex, lesion location) from the PAD-UPES-20 dataset. The **intra- and inter-modality attention mechanisms** of their model achieved an average accuracy of **76.8%**, demonstrating the utility of metadata. However, limitations such as low accuracy, a small dataset, and restricted lesion diversity highlight challenges in generalizability.

Restrepo et al. [23] explored the use of **vector embeddings** extracted from foundation models for **multimodal data fusion** in **low-resource settings**, comparing it with traditional raw data processing. Their study investigated **unimodal embeddings** (DINO v2 for images, LLAMA 2 for text), **VLM embeddings** (CLIP), and raw data fine-tuning using BERT and ViT, evaluated on **BRSET**, **HAM10000**, and **SatelliteBench** datasets. Using **early and late fusion techniques**, the embedding approach reduced **computational costs** while maintaining high accuracy (**0.987**) and F1-score (**0.944**) on BRSET. The accuracy achieved on the **HAM10000** dataset was comparatively lower, which can be attributed to **domain-specific challenges** inherent to dermatology. These challenges include subtle variations in skin lesion features and a potential mismatch between the training data of foundation models and the specific domain of the HAM10000 dataset. However, the use of embedding alignment techniques contributed to improved performance. Despite these advancements, the study's reliance on specific datasets and the training data of foundation models may limit its generalizability. Additionally, the evaluation conducted in a low-resource CPU-only environment may not accurately reflect real-world performance, which typically leverages GPU capabilities for enhanced computational efficiency.

To summarize, despite notable progress having been made on both single-modality and multimodal skin cancer classification, some of the shortcomings still remain. Numerous current models are computationally intensive, rendering them impractical to use in real-world applications, particularly where there are constraints on resources. These models also have difficulty generalizing across varied datasets, and their performance becomes suboptimal when they have to handle other unknown classes or metadata. Additionally, while multimodal systems have shown improved accuracy, they still suffer from limitations like overfitting and inadequate training and inference efficiency.

Our method is designed to address these limitations by building on fusion methods that integrate skin images and clinical data, improving classification performance while being efficient. We focus on minimizing computational expenses and storage needs, rendering the system more implementable in resource-constrained environments. In addition, to enhance interpretability, we incorporate the Grad-CAM method so that feature relevance can be better evaluated and the system can be made more clinically usable. Finally, our method aims to deliver a strong, efficient, and interpretable solution for skin lesion classification in real-world, resource-limited settings. Table 1 consolidates the above papers

3 MATERIALS AND METHODS

3.1 Dataset

The data we have used for our classification problem is the
textbfHAM10000 dataset
citehampaper. HAM10000, which is ”Human Against Machine,” is a dataset containing multi-source dermatoscopic images of pigmented lesions. The final dataset contains
textbf10,015 dermatoscopic images acquired over 20 years from two main sources: the Department of Dermatology at the Medical University of Vienna, Austria, and a skin cancer practice in Queensland, Australia. The data consists of **7 diagnostic classes**, representing the most frequent pigmented skin lesions. Table 2 gives the distribution of the classes.

The dataset also provides metadata for each patient, including clinical information such as age, gender, and lesion location. Using an `image_id` column, the metadata of the patient can be linked to its corresponding lesion image. This metadata is outlined in Table 3.

3.2 Preprocessing Pipeline

The preprocessing pipeline for the HAM10000 dataset involved several steps to ensure high-quality input data for our multimodal fusion approach. These steps are outlined below:

3.2.1 Class Balancing

The HAM10000 dataset exhibits significant class imbalance, with some classes having as few as 115 images (e.g., Dermatofibroma) and others having over 6,000 images (e.g., Melanocytic Nevus). To address this, we applied the following data augmentation techniques:

- **Replication**: Duplicated samples from minority classes.
- **Jittering**: Added random noise to pixel values.
- **Random Sampling**: Randomly selected subsets of majority classes.
- **Geometric Transformations**: Applied horizontal and vertical flips, rotations, and scaling.

After augmentation, each class contained 6,000 images, resulting in a balanced dataset of 42,000 images.

3.2.2 Image Preprocessing

Each dermatoscopic image underwent the following preprocessing steps:

- **Resizing**: Images were resized to a uniform resolution of 256x256 pixels.
- **Normalization**: Pixel values were normalized to the range [0, 1] by dividing by 255.

3.2.3 Metadata Preprocessing

- **Handling Missing Values**: Missing values in metadata (e.g., age, gender) were imputed using the median for numerical features and the mode for categorical features. The median was chosen for numerical features like age because it is robust to outliers, ensuring that extreme values do not skew the imputation. For categorical features like gender, the mode was selected as it represents the most frequent category, preserving the distribution of the data. Experiments showed that using the median for numerical features reduced the impact of outliers by 15% compared to the mean, while the mode for categorical features maintained the original class distribution with 98% accuracy.
- **Normalization**: Numerical features (e.g., age) were normalized to have zero mean and unit variance.
- **Encoding**: Categorical features (e.g., gender, lesion location) were one-hot encoded.

3.2.4 Data Alignment

To ensure proper alignment between images and metadata, we used the `image_id` column to map each image to its corresponding clinical metadata. This ensured that the multimodal fusion model received correctly paired inputs during training and evaluation. Following this the data was split into three sets:

- **Training Set** : 70%
- **Testing Set** : 15%
- **Validation Set** : 15%

The flow has been visualized in Figure 1.

3.3 Model Pipeline Process

The multimodal fusion pipeline consists of three main components: (1) a custom Weighted ResNet for extracting features from dermatoscopic images, (2) a Clinical CNN for processing clinical metadata, and (3) a fusion module that combines the extracted features for final classification. The pipeline operates as follows:

- **Input**: Dermatoscopic images and clinical metadata are preprocessed and fed into the pipeline.
- **Feature Extraction**: The Weighted ResNet processes the images, while the Clinical CNN processes the metadata.
- **Fusion**: The extracted features are combined using one of several fusion techniques that are further discussed in the paper
- **Classification**: The fused feature vector is passed through a Feed-forward Neural Network (FFN) to predict the skin lesion class.

The details of all networks has been provided in the architecture section and the pipeline has been illustrated in Figure 2

3.4 Architecture

3.4.1 Clinical CNN

Convolutional Neural Networks (CNNs) are a class of deep learning models which are designed to process grid-like data, such as images, by leveraging convolutional layers to extract spatial relationships of features [31]. CNNs consist of multiple layers, including convolutional layers, pooling layers, and fully connected layers, which work together to capture complex information including spatial representations of features. These models have been widely adopted in computer vision tasks due to their ability to capture local patterns and spatial dependencies. [32].

In this work, the CNN architecture is adapted to process tabular patient metadata, such as age, gender, and lesion location. While CNNs are traditionally used for image data, their ability to learn hierarchical features makes them suitable for structured data as well. By treating the metadata as a 1D feature vector, we apply 1D convolutional layers to extract meaningful patterns and relationships from the data.

The Clinical CNN is designed to process tabular patient metadata, by transforming the input into a compact feature vector. The model consists of two fully connected layers: the first layer maps the input metadata to a 128-dimensional hidden space using a ReLU activation function, while the second layer further processes the features into a 256-dimensional representation. This is a simple architecture aimed not only to extract relevant features from the patient metadata but also to be computationally efficient. The architecture has been displayed in figure 3

3.4.2 Weighted ResNet

A ResNet or Residual Networks are a CNN-Based Architecture derived from AlexNet first developed in 2015 for the ImageNet 2015 competition that aimed to address the degradation problem in deep neural networks. ResNets work by breaking down a deep neural network into smaller segments known as residual blocks which are then connected through skip or residual connections. The layers learn residual functions with reference to layer inputs. This allows gradients to flow more easily during backpropagation. Consider a residual block with two convolutional layers. Let W_1 and W_2 represent the weights of the convolutional layers, and σ denote the activation function (e.g., ReLU). The output of the residual block can be expressed as:

$$F(x) = W_2 \cdot \sigma(W_1 \cdot x) \tag{1}$$

The final output of the block is:

$$y = F(x) + x \tag{2}$$

This formulation allows the network to learn residual mappings, which are often easier to optimize than the original unreferenced mappings. [33] [34]

DermiResNet extends the traditional ResNet architecture by introducing learnable weights for the residual connections. Instead of simply adding the input x to the output of the residual block, a learnable weight α is applied:

$$y = F(x) + \alpha \cdot x \tag{3}$$

Here, α is a learnable parameter that controls the contribution of the skip connection. This allows the network to dynamically adjust the importance of the residual path during training. [33], [35]

The model, `DermiResNet`, is a custom implementation of a Residual Network. It consists of a series of convolutional and residual blocks, with learnable weights applied to the skip connections. The architecture begins with an initial convolutional block (`conv1`) that reduces spatial dimensions and increases channels to 64. This is followed by four main stages, each comprising a convolutional block and a residual block. The convolutional blocks (`conv2`, `conv3`, `conv4`, `conv5`) progressively increase the number of channels (64, 128, 256, 512) while halving spatial dimensions using a stride of 2. Each residual block (`res1`, `res2`, `res3`, `res4`) consists of two convolutional layers with batch normalization and LeakyReLU activations. The output of each residual block is added to its input, weighted by a learnable parameter `res_weights`, to form the skip connection. Dropout is applied in `res3` and `res4` for regularization. The model concludes with a fully connected module, which includes adaptive average pooling, flattening, and two fully connected layers. A softmax function is applied to produce a 512 feature sample space. This architecture allows the network to dynamically adjust the contribution of skip connections during training, improving feature learning and optimization. The model is shown in figure 4 and figure 5.

3.4.3 Fusion Block

In this layer, we fuse the extracted features from the DermiResNet (image features) and the Clinical CNN (clinical metadata features). The fusion process combines these multimodal features into a unified representation, which is then passed to a simple classifier for final prediction. The classifier consists of fully connected layers with ReLU activation functions, reducing the fused feature space to 7 output classes corresponding to the skin lesion types.

We explore several fusion techniques to combine the features effectively, including:

- Simple Concatenation
- Weighted Concatenation
- Hadamard Product
- Tensor Fusion
- Bilinear Fusion
- Gated Fusion
- Self-Attention
- Cross-Attention

Detailed descriptions of these fusion techniques, including their mathematical formulations and implementation, are provided in Section **??**

3.5 Fusion Details

Fusion refers to the process of combining information from multiple sources or modalities to create a unified representation. This is particularly useful in multimodal learning, where data from different domains (e.g., text, images, audio) are integrated to improve model performance. In this model the features recieved from the DermiResNet and those extracted from the Clinical CNN were fusion. In this section each fusion technique used has been described. The outcomes received from each technique have been compared and contrasted

3.5.1 Simple Concatenation

Simple concatenation involves merging features from different modalities by **stacking** them along a specific dimension.

Given two feature vectors $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$, the concatenated feature vector $\mathbf{z} \in \mathbb{R}^{n+m}$ is:

$$\mathbf{z} = [\mathbf{x}; \mathbf{y}] \tag{4}$$

Equation (4) shows the simple concatenation of two feature vectors.

3.5.2 Weighted Concatenation

Weighted Concatenation is a simple evolution from simple concatenation where instead of simply stacking the features from different modalities, each feature is assigned a **weight** before concatenating. This weight can be fixed or it can be **learnable**. This will allow the model to assign different importance to each modality, giving priority to one feature over another. Weighted concatenation is useful in scenarios where one modality is more reliable or informative than the other [36, 37]. In our approach, we assigned learnable weights.

Given weights w_1 and w_2, the weighted concatenated feature vector $\mathbf{z}$ is:

$$\mathbf{z} = [w_1\mathbf{x}; w_2\mathbf{y}] \tag{5}$$

Equation (5) defines the weighted concatenation of two feature vectors.

3.5.3 Hadamard Product Fusion

The Hadamard product, or **element-wise multiplication**, is a fusion technique that combines features by multiplying corresponding elements of feature vectors. This method emphasizes the interaction between modalities at a fine-grained level, making it suitable for tasks where local feature interactions are important [38].

For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the Hadamard product $\mathbf{z} \in \mathbb{R}^n$ is:

$$\mathbf{z} = \mathbf{x} \odot \mathbf{y} \tag{6}$$

Equation (6) represents the Hadamard product of two feature vectors.

3.5.4 Tensor Fusion

Tensor fusion creates a higher-dimensional representation by computing the **outer product** of feature vectors from different modalities. While this method can capture feature-rich interaction between different modalities, it can be computationally expensive due to the high dimensionality of the resulting tensor [39].

For $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$, the tensor $\mathbf{Z} \in \mathbb{R}^{n \times m}$ is:

$$\mathbf{Z} = \mathbf{x} \otimes \mathbf{y} \tag{7}$$

Equation (7) shows the tensor fusion of two feature vectors.

3.5.5 Bilinear Fusion

Bilinear fusion combines features using a **bilinear transformation**, which explicitly **models pairwise interactions** between modalities. This method is particularly effective for tasks where the relationship between modalities is nonlinear and complex [40].

For $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$, and a learnable matrix $\mathbf{W} \in \mathbb{R}^{n \times m}$, the bilinear fusion output $\mathbf{z} \in \mathbb{R}^k$ is:

$$\mathbf{z} = \mathbf{x}^\top \mathbf{W} \mathbf{y} \tag{8}$$

Equation (8) defines the bilinear fusion of two feature vectors.

3.5.6 Gated Fusion

Gated fusion employs a **gating mechanism** to **dynamically control** the contribution of each modality based on the input data. A gating mechanism is essentially a function that determines how much of each modality's information should be retained or suppressed. In our model, we use the **sigmoid** function as the gating mechanism, which outputs values between 0 and 1. These values act as ”gates” that modulate the influence of each modality. For example, if the gate value for a modality is close to 1, the modality's features are heavily utilized, whereas a value close to 0 suppresses its contribution. Other commonly used gating mechanisms are **tanh**, **ReLU** and **softmax**. The key difference between gated fusion and the previously mentioned weighted concatenation is that gated fusion allows the model to dynamically adjust the importance of each feature for every input, whereas in weighted concatenation, the importance remains fixed for all inputs once trained [41].

For $\mathbf{x}$ and $\mathbf{y}$, the gated fusion output $\mathbf{z}$ is:

$$\mathbf{z} = \sigma(\mathbf{W}_x \mathbf{x} + \mathbf{W}_y \mathbf{y}) \odot \mathbf{x} + (1 - \sigma(\mathbf{W}_x \mathbf{x} + \mathbf{W}_y \mathbf{y})) \odot \mathbf{y} \tag{9}$$

Equation (9) represents the gated fusion of two feature vectors, where σ is the sigmoid function.

3.5.7 Self-Attention Fusion

Self-attention computes attention weights within a single modality to focus on the most relevant features. This mechanism enables a model to focus on different parts of the **same input sequence** when making predictions. It allows the model to compute relationships between all elements in the sequence, creating a context-aware representation. For example, in natural language processing, self-attention helps the model understand which words in a sentence are most relevant to each other, even if they are far apart [42].

For an input sequence $\mathbf{X} \in \mathbb{R}^{n \times d}$ (where n is the sequence length and d is the feature dimension), self-attention computes three learned transformations:

- **Query (Q)**: Represents the current element being processed.
- **Key (K)**: Represents the elements to be compared with the query.
- **Value (V)**: Represents the information to be aggregated.

The self-attention output $\mathbf{Z}$ is computed as:

$$\mathbf{Z} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V} \tag{10}$$

Equation (10) shows the self-attention mechanism, where $\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}$ computes the similarity between queries and keys, and the softmax function normalizes these similarities into attention weights.

3.5.8 Cross-Attention Fusion

Cross-attention computes attention weights between two modalities to align and combine their features. Cross-attention extends the idea of self-attention by computing relationships between elements of **two different sequences**. It allows one sequence to ”attend” to another, enabling the model to align and integrate information across modalities. For example, in multimodal tasks, cross-attention can help align words in a text description with regions in an image. This method is particularly useful for tasks where the relationship between modalities is asymmetric or hierarchical, such as aligning text with image regions [43].

For two input sequences $\mathbf{X} \in \mathbb{R}^{n \times d}$ and $\mathbf{Y} \in \mathbb{R}^{m \times d}$, cross-attention computes:

- **Query ($\mathbf{Q}_x$)**: Derived from the first sequence $\mathbf{X}$.
- **Key ($\mathbf{K}_y$)** and **Value ($\mathbf{V}_y$)**: Derived from the second sequence $\mathbf{Y}$.

The cross-attention output $\mathbf{Z}$ is computed as:

$$\mathbf{Z} = \text{softmax}\left(\frac{\mathbf{Q}_x\mathbf{K}_y^\top}{\sqrt{d}}\right)\mathbf{V}_y \tag{11}$$

Equation (11) defines the cross-attention mechanism, where $\frac{\mathbf{Q}_x\mathbf{K}_y^\top}{\sqrt{d}}$ computes the similarity between queries from $\mathbf{X}$ and keys from $\mathbf{Y}$, and the softmax function produces attention weights.

3.6 Experimental Set-Up

3.6.1 Hardware and Software Configuration

The hardware and software specifications used for the experiments are summarized in Table 4.

3.6.2 Training Configuration

The specific training configuration, which outlines hyperparameters and other details, is documented in Table 5. Each model was trained for an estimated duration of two hours.

3.7 Evaluation Metrics

The performance of the model was evaluated using Accuracy, Precision, Recall, F1-Score and AUC-ROC

4 RESULT ANALYSIS AND DISCUSSION

4.1 Ablation Studies

To assess the contribution of each modality to the overall model performance, we first evaluated the individual models trained separately on clinical metadata and dermatoscopic images. The results of these experiments are summarized in Table 6.

The Clinical CNN achieved an accuracy of **77.0%**, indicating that clinical metadata alone offers moderate predictive capability. However, the DermiResNet, trained exclusively on dermatoscopic images, achieved a significantly higher accuracy of **92.0%**, showcasing the superior discriminative power of visual data for skin lesion classification. This result aligns with the diagnostic process commonly employed by dermatologists, where visual inspection of skin lesions is typically prioritized over metadata analysis for accurate classification [44].

4.2 Multimodal Fusion Performance

We evaluated the performance of various fusion techniques, including simple concatenation, weighted concatenation, Hadamard product, tensor fusion, bilinear fusion, gated fusion, self-attention, and cross-attention. The results are summarized in Table 7.

The multimodal fusion techniques reveal profound insights into computational medical diagnostics, with the **Cross-Attention** and **Hadamard Product** methods demonstrating exceptional performance, approaching **99% accuracy**. The remarkable success of Cross-Attention (**98.86% accuracy**) can be attributed to its sophisticated mechanism of dynamically weighting and aligning features from clinical metadata and dermatoscopic images. By computing attention scores between modalities, Cross-Attention effectively mimics the nuanced diagnostic reasoning of experienced clinicians, adaptively focusing on the most discriminative information for each input. For instance, in cases where visual features are ambiguous (e.g., lesions with similar appearances), Cross-Attention leverages clinical metadata (e.g., age, lesion location) to refine predictions, bridging the gap between visual and contextual information. This ability to model fine-grained, adaptive interactions between modalities makes Cross-Attention particularly effective for complex tasks like skin lesion diagnosis, where both visual patterns and patient-specific factors play a crucial role. [22]

Similarly, the **Hadamard Product** (**98.85% accuracy**) excels by capturing localized relationships between modalities through element-wise multiplication of feature vectors. This method is particularly effective at highlighting subtle interactions between visual and clinical features, such as the correlation between lesion texture and patient age. However, unlike Cross-Attention, the Hadamard Product lacks the dynamic adaptability to adjust feature importance based on the input, which may limit its performance in more complex or ambiguous cases.

In contrast, the relatively poor performance of **Gated Fusion** (**93.0%**) and **Self-Attention** (**92.70%**) suggests that more complex neural architectures are not always superior. These techniques may suffer from increased model complexity, potential overfitting, and sensitivity to hyperparameter configurations, especially when working with limited training data. For instance, Gated Fusion's dependency on gated mechanisms learned can bring in extra parameters hard to optimize, and Self-Attention's emphasis on intra-modality interactions might neglect key inter-modality relationships. The significant performance disparity among these approaches highlights the utmost significance of multimodal fusion technique design, wherein the capability of incorporating heterogeneous sources of data with a sense of intelligence can substantially contribute to diagnostic accuracy. .

notably, less complex methods such as **Weighted Concatenation** (**97.15%**) show that occasionally a simpler strategy can get close to state-of-the-art performance. Through linear weighting of features from the two modalities, Weighted Concatenation presents a low-cost and stable fusion method with less opportunity for overfitting.

This shows that algorithmic complexity isn't always directly associated with prediction capability in machine learning models for medical image classification. Rather, the selection of fusion method should be determined by the particular properties of the dataset and clinical setting, trading off performance, interpretability, and computational expense.

These findings have significant implications for the design of multimodal diagnostic systems. While simpler methods like Weighted Concatenation offer a practical trade-off between accuracy and computational cost, advanced techniques like Cross-Attention are better suited for scenarios where maximizing performance is critical, particularly when the interplay between visual and clinical features is complex and nuanced.

Ultimately, the success of these fusion techniques lies in their ability to intelligently integrate visual and clinical data, mirroring the holistic diagnostic approach of human clinicians and paving the way for more reliable and interpretable AI-driven healthcare solutions

4.3 Confusion Matrix Analysis

To further analyze the performance of the best-performing fusion model (Cross-Attention), we present its confusion matrix in Table 8. The matrix shows the number of correct and incorrect predictions for each class.

The confusion matrix highlights the performance of the model across various skin cancer classes. Notably, the diagonal entries, which represent correctly classified instances, dominate, demonstrating strong overall performance.

The model achieves near-perfect classification for classes like **Dermatofibroma (df)**, **Vascular lesions (vasc)**, and **Actinic keratoses (akiec)**, with minimal or no misclassifications. This suggests that these classes are well-represented in the training data and exhibit distinct features, allowing the model to identify them with high confidence.

However, there are minor misclassifications, particularly between similar-looking lesion types such as **Benign keratosis (bkl)** and **Melanoma (mel)** or **Melanocytic nevi (nv)**. For instance, 7 cases of **bkl** are misclassified as **mel**, and 5 as **akiec**, indicating some overlap in visual or clinical features. Similarly, 4 cases of **nv** are mistaken for **mel**, which is expected given their subtle differences and shared features in certain instances.

The small number of misclassifications in **Basal cell carcinoma (bcc)** and **Melanoma (mel)** classes suggests the model handles malignant lesions well but still requires improvements to reduce errors in high-stakes scenarios.

Overall, the model demonstrates high classification accuracy with room for improvement in distinguishing between lesion types with overlapping visual or clinical characteristics.

4.4 ROC-AUC Curve

The ROC-AUC scores for different classes are presented in the figure below. The model achieved a score of 0.99 for melanoma, 0.98 for nevi, and 1.0 for the remaining classes, as shown in the ROC curve (Figure 6).

These high AUC values suggest that the model is capable of effectively distinguishing between the classes. The results indicate that there is no sign of underfitting or overfitting, with the model generalizing well and avoiding excessive fitting to noise in the training data.

4.5 Explainability

Explainability is a foundation of applying machine learning models in healthcare. Although deep learning models tend to exhibit state-of-the-art performance, their ”black-box” nature is highly problematic in the healthcare domain. Physicians and medical experts need interpretable models for them to comprehend the rationale of predictions, making diagnoses not only correct but also clinically reasonable. Lack of transparency can translate to mistrust, preventing the implementation of AI systems into actual healthcare workflows.

Black-box models that give no clue about their reasoning are especially troubling in high-risk areas such as dermatology. For example, a model can be highly accurate by using spurious correlations or data artifacts instead of clinically significant features. This can result in disastrous failures when the model is applied to heterogeneous or novel situations. Explainability closes this gap by illuminating how a model comes to a decision, allowing doctors to verify its reasoning and spot potential errors or biases. [45]

We, in our research, emphasize explainability so that our multimodal fusion model is not just precise but also reliable and comprehensible. Through the use of visual explanations (Grad-CAM) coupled with relevance analysis of clinical features, we present an integrated understanding of the decision-making process of the model.

4.5.1 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) is a powerful technique for visualizing the regions of an image that are most influential in a model's decision. It extends the Class Activation Mapping (CAM) approach by using gradient information to weight the importance of feature maps, making it applicable to a wider range of architectures, including those without global average pooling layers.

Mathematical Formulation: Let A^k be the activation map of the k-th channel in the target convolutional layer, and let y^c be the score for class c. The weight α_k^c for the k-th channel is computed as the global average of the gradients of y^c with respect to A^k:

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

where Z is the number of pixels in the activation map. These weights capture the importance of each feature map for the target class.

The Grad-CAM heatmap L^c is then obtained by a weighted combination of the activation maps, followed by a ReLU function:

$$L^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right).$$

The ReLU function ensures that only positive influences are considered, as negative values are not relevant for the target class. [46]

Grad-CAM highlights the regions of the image that the model deems most important for its prediction. For example, in dermatoscopic images, these regions might correspond to lesion borders, texture, or color variations. By visualizing these regions, Grad-CAM provides a window into the model's "thought process," enabling doctors to verify that the AI system is focusing on clinically meaningful features. [47]

4.5.2 Analysis of Grad-CAM Results

To gain insights into the decision-making process of our best-performing model, which employs a cross-attention fusion mechanism, we applied Grad-CAM visualizations. Grad-CAM enables the interpretation of the model's predictions by highlighting regions within dermatoscopic images that contribute most significantly to the output. This form of post-hoc explainability is particularly important in medical applications, as it provides clinicians with a means to verify that the model bases its predictions on clinically relevant features.

The cross-attention fusion mechanism enhances the interpretability of the model by dynamically aligning and integrating features from different modalities or scales. Unlike simpler fusion techniques, cross-attention computes attention weights between modalities, allowing the model to focus on the most relevant interactions. This results in feature representations that are both context-aware and clinically meaningful, as evidenced by the Grad-CAM visualizations. Compared to other architectures, the cross-attention model produces heatmaps that are sharper, more precise, and aligned with diagnostic criteria.

Figure 7 provides an illustrative example of a Grad-CAM heatmap overlaid on an input dermatoscopic image. The left panel displays the original image, which corresponds to a lesion diagnosed as melanoma. The right panel shows the Grad-CAM heatmap, with red regions indicating areas of high importance and blue regions denoting areas of low importance. The model predicts the lesion as melanoma with a confidence score of 0.98, focusing primarily on the lesion's irregular borders and regions of heterogeneous pigmentation. These features are clinically significant, as melanomas are characterized by asymmetry, border irregularity, and color variation.

In this specific example, the heatmap demonstrates that the model successfully identifies features associated with malignancy while ignoring irrelevant artifacts, such as hair strands and uniform skin areas. The model's ability to focus on clinically relevant features is a direct outcome of the cross-attention fusion mechanism, which dynamically aligns and weights features from different modalities. By doing so, the model achieves a higher level of alignment with dermatological practices, where lesion borders and internal variations are critical for diagnosis.

As discussed earlier, the model occasionally misclassifies Benign Keratosis (bkl) as Melanoma (mel) due to overlapping visual characteristics. One contributing factor is that both lesion types can exhibit irregular pigmentation, asymmetric structures, and varying border definitions, which are also key diagnostic markers for melanoma. This confusion is particularly evident in cases where keratosis presents with darker pigmentation and irregular borders, mimicking features of malignant lesions.

Figure 8 illustrates such a misclassification, where a Benign Keratosis lesion has been incorrectly predicted as Melanocytic (mel) with a confidence score of 0.46. The left panel shows the original image, while the right panel presents the corresponding Grad-CAM heatmap, highlighting the regions that influenced the model's decision.

From the heatmap, it is evident that the model assigns high importance (red/yellow regions) to darker pigmented areas and irregular structures within the lesion. This suggests that the model's decision boundary between benign and malignant lesions is influenced primarily by pigmentation and border irregularity, which are not always exclusive to melanoma.

Beyond this example, Grad-CAM visualizations across a wide range of test cases reveal consistent patterns in the model's behavior. The model often prioritizes irregular lesion borders and regions of color variation, which are crucial for distinguishing malignant lesions from benign ones. In cases of basal cell carcinoma, the heatmaps highlight central regions with ulceration or shiny surfaces, further reinforcing the model's alignment with clinical indicators. This interpretability is invaluable for real-world deployment, as it allows clinicians to confirm that the model's focus areas correspond to meaningful diagnostic features.

The insights provided by Grad-CAM are directly correlated with the model's strong quantitative performance. For example, the regions identified by the heatmaps frequently align with features responsible for the model's high sensitivity and specificity, particularly in challenging cases of melanoma and basal cell carcinoma. This combination of performance metrics and visual interpretability underscores the potential of the cross-attention fusion-based model as a trustworthy diagnostic aid.

4.5.3 Clinical Feature Relevance

To assess the instance-specific relevance of clinical features, we employed an attribution-based approach using Integrated Gradients (IG) [48]. IG is a path-based attribution method that quantifies feature importance by computing the integral of gradients along an interpolation path from a baseline input to the actual input.

Given an input clinical feature vector $x \in \mathbb{R}^d$ and a baseline vector $x' \in \mathbb{R}^d$, the integrated gradient for the i^{th} feature is computed as:

$$IG_i(x) = (x_i - x'i) \times \int \alpha = 0^1 \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha, \tag{12}$$

where $f(\cdot)$ represents the model's prediction score for the target class. The integral is approximated using a summation over discrete steps:

$$IG_i(x) \approx (x_i - x'i) \times \frac{1}{S} \sum s = 1^S \frac{\partial f(x' + \frac{s}{S}(x - x'))}{\partial x_i}, \tag{13}$$

where S is the number of interpolation steps.

For each instance, we set the baseline x' as a zero vector, representing the absence of clinical features. The integrated gradients were computed over $S = 50$ steps, and the absolute values of the attributions were taken as feature importance scores. These scores were then normalized to the range $[0, 1]$ to facilitate interpretability.

The resultant feature relevance scores highlight the contribution of individual clinical attributes to the model's prediction. Features with higher attributions indicate stronger influence on the classification decision, thereby providing insights into the model's reliance on clinical metadata.

4.5.4 Clinical Feature Relevance Analysis

For a specific instance of **Melanocytic Nevus**, we analyzed the relative importance of clinical features using our best-performing Cross-Attention fusion model. As shown in Table 9, the top features with high importance include **localization** (scalp, foot, neck) and **diagnostic methods** (consensus, follow-up, confocal). These features likely play a significant role due to the distinct characteristics of lesions in these locations and the reliability of the diagnostic methods. Features with moderate importance, such as localization (face, genital) and diagnosis type (histopathology), contribute to predictions but are less critical. Interestingly, **age** and certain locations (e.g., ear, lower extremity) show low importance, suggesting they have minimal influence on the model's predictions for this instance. The balanced impact of **sex** (both male and female) indicates that while it influences outcomes, it is not among the strongest predictors. Overall, localization emerges as the most relevant feature, aligning with real-world clinical practice where lesion location is a key diagnostic factor.

Across various lesion types, we observed that **localization** consistently emerged as the most important clinical feature, underscoring its critical role in skin lesion diagnosis. This aligns with real-world clinical practice, where the anatomical location of a lesion is a key diagnostic factor due to its correlation with sun exposure, skin type, and lesion characteristics.

In cases of **Melanoma**, **localization** (e.g., back, face) was the top predictor, reflecting the higher prevalence of malignant lesions in sun-exposed areas. Additionally, **age** and **gender** showed moderate influence, consistent with their established roles as risk factors for melanoma. Older patients and males exhibited a higher likelihood of malignancy, further validating the model's alignment with epidemiological trends.

For **Basal Cell Carcinoma (BCC)**, the model again prioritized **localization** (e.g., face, neck), as these areas are most susceptible to UV damage. Diagnostic methods such as **histopathology** also played a significant role, as BCC is often confirmed through biopsy. Interestingly, **age** showed a stronger influence for BCC compared to other lesion types, likely due to the cumulative effect of UV exposure over time.

In cases of **Actinic Keratosis (AKIEC)**, **localization** (e.g., face, scalp) remained the most important feature, as AKIEC lesions are strongly associated with chronic sun exposure. The model also highlighted the importance of **diagnostic methods** (e.g., follow-up, confocal microscopy), reflecting the need for repeated evaluations to monitor these precancerous lesions.

For **Dermatofibroma (DF)**, a benign lesion, **localization** (e.g., lower extremities) was again the dominant feature, while **age** and **gender** had minimal influence. This suggests that the visual appearance and location of DF lesions are more critical for diagnosis than demographic factors.

Finally, in cases of **Vascular Lesions (VASC)**, **localization** (e.g., face, trunk) was the most influential feature, as these lesions often appear in specific anatomical regions. The model also relied heavily on **dermoscopic examination**, highlighting the importance of visual data for diagnosing vascular lesions.

5 CONCLUSION

The effectiveness of multimodal fusion methods in improving skin lesion classification performance is well illustrated through this study. By combining clinical metadata with dermatoscopic images, the model reached state-of-the-art accuracy, with the **Cross-Attention** and **Hadamard Product** methods reaching almost **99% accuracy**. Importantly, these accuracies were reached with a basic laptop setup, without the necessity for specialized NPUs or workstations, making the proposed method efficient and accessible. This is especially important for resource-limited settings, where sophisticated computational facilities might not be easily accessible.

The strength of these fusion methods rests in their capacity to merge complementary information from visual and clinical data, emulating the comprehensive diagnostic reasoning that seasoned clinicians often exhibit. The
textbfCross-Attention method, specifically, performed exceptionally well by dynamically correlating and weighting features from both modalities to allow the model to concentrate on the most discriminative information per input. Likewise, the
textbfHadamard Product also performed well by appropriating local modality relationships through element-wise multiplication of feature vectors.

These results highlight the need for carefully choosing fusion techniques depending on the specific characteristics of the dataset and clinical context. Though more sophisticated methods like Cross-Attention present better performance in challenging scenarios, simpler approaches such as **Weighted Concatenation** offer a useful trade-off between accuracy and computational expense.

5.1 Limitations and Future Directions

Although the findings of this study highlight the capability of multimodal fusion to improve skin lesion classification, it is essential to recognize various limitations requiring future work and improvement.

- **Computational Resource Requirements**: While the adopted methodologies illustrate feasibility on common computing hardware, the inherent computational intensity of sophisticated fusion methods, specifically tensor fusion and cross-attention, poses a significant computational burden. Future studies should focus on developing and utilizing optimization techniques. Model pruning, quantization, and knowledge distillation are some techniques that need to be explored to reduce computational overhead without compromising diagnostic performance.
- **Generalizability Across Diverse Populations**: The use of the HAM10000 dataset, as comprehensive as it is, may not perfectly capture the heterogeneity of skin lesions observed in real-world clinical practice. Therefore, the generalizability of the model to diverse patient populations remains an important challenge. Future research should include external validation by evaluating performance on datasets such as ISIC and PH2. Additionally, expanding training data with a broader range of clinical and demographic variables is crucial to enhancing the model's robustness and validity.
- **Integration of Extensive Clinical Data**: The predictive capability of the present model is constrained by the limited range of clinical features available in the HAM10000 dataset. To address this, future research should focus on integrating more comprehensive clinical data. This includes patient medical

histories, genetic predisposition, and laboratory test results, which collectively contribute to a more holistic diagnostic analysis.

- **Improving Model Interpretability for Clinical Trust**: Clinical adoption of AI-based diagnostic tools is contingent on their interpretability. Although Grad-CAM provides a visual interpretation of feature importance, a deeper understanding of the model's decision-making process is necessary. Future studies should explore advanced Explainable AI (XAI) techniques, such as combining Grad-CAM with clinical feature relevance analysis or developing hybrid models that provide both visual and textual explanations.
- **Refinement of Fusion Methodologies**: The success of multimodal fusion depends on the selection and optimization of appropriate techniques. Future research should explore adaptive fusion methods that dynamically adjust based on input features. Additionally, investigating ensemble fusion techniques that leverage the strengths of multiple fusion strategies could lead to significant improvements in diagnostic accuracy.
- **Validation in Real-World Clinical Environments**: To assess the practical effectiveness of the proposed system, rigorous validation in real-world clinical settings is essential. Future studies should emphasize real-time deployment of the model in diagnostic workflows, ensuring close collaboration with dermatologists and healthcare professionals to address implementation challenges and optimize the system based on real-world feedback.
- **Handling Class Imbalance and Rare Phenotypes**: The misclassification of rare transitions, such as melanoma being classified as nevus or benign keratosis, underscores the need for better handling of class imbalances and subtle feature variations. Future research should explore techniques like oversampling, undersampling, and synthetic data generation to mitigate class imbalance effects. Additionally, feature-aware loss functions and attention mechanisms that emphasize subtle but clinically significant features could enhance the model's ability to distinguish between highly similar lesion types.
- **Improving Preprocessing Resilience**: The tendency of the model to focus on non-essential regions, such as hair strands or glossy skin, highlights the need for more robust preprocessing techniques. Enhancing hair removal algorithms and contrast normalization strategies is crucial to eliminating distractions and improving the model's reliability in assessing key lesion features.

This research lays a strong foundation for applying multimodal fusion in skin lesion classification. By addressing the identified limitations and exploring the proposed future directions, we can advance the development of precise, efficient, and clinically viable AI-driven diagnostic systems, ultimately leading to improved outcomes for patients with dermatological conditions.

CONFLICT OF INTEREST STATEMENT

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors have no affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter discussed in this manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization, A.D. and N.P.S.; methodology, A.D. and V.A.; software, A.D.; validation, A.D., V.A. and N.P.S.; formal analysis, V.A.; investigation, A.D. and V.A.; resources, N.P.S.; data curation, A.D.;

writing–original draft preparation, A.D. and V.A.; writing–review and editing, N.P.S.; visualization, V.A.; supervision, N.P.S.; project administration, N.P.S. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank the Department of Information and Communication Technology, Manipal Institute of Technology for providing the necessary research infrastructure. We also acknowledge Cryptonite Student Project's Research Division for their valuable insights during the initial phases of this study.

DATA AVAILABILITY STATEMENT

The dataset analyzed during this study is publicly available in the Skin Cancer MNIST: HAM10000 repository on Kaggle. The dataset is readily available in kaggle as Skin Cancer MNIST [49]. This dataset consists of dermatoscopic images from different populations, acquired and stored by different modalities. The images were labeled according to established clinical classification methods.

REFERENCES

[1]Nazim Hasan, Arshad Nadaf, Mohd Imran, Umar Jiba, Asim Sheikh, Wejdan H Almalki, Saleh S Almujri, Yousuf H Mohammed, Prashant Kesharwani, and Farhan J Ahmad. Skin cancer: understanding the journey of transformation from conventional to advanced treatment approaches. *Molecular Cancer*, 22(1):168, 2023.

[2]Kavita, JS Thakur, and T Narang. The burden of skin diseases in india: Global burden of disease study 2017. *Indian Journal of Dermatology, Venereology and Leprology*, 89(3):421–425, 2023.

[3]Howard W. Rogers, Martin A. Weinstock, Steven R. Feldman, and Brett M. Coldiron. Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the us population, 2012. *JAMA Dermatology*, 151(10):1081–1086, 10 2015.

[4]A. F. Jerant, J. T. Johnson, C. D. Sheridan, and T. J. Caffrey. Early detection and treatment of skin cancer. *Am Fam Physician*, 62(2):357–68, 375–6, 381–2, Jul 2000.

[5]Lois J. Loescher, Monika Janda, H. Peter Soyer, Kimberly Shea, and Clara Curiel-Lewandrowski. Advances in skin cancer early detection and diagnosis. *Seminars in Oncology Nursing*, 29(3):170–181, 2013. Skin Cancer.

[6]Gunjan Rajput, Shashank Agrawal, Gopal Raut, and Santosh Kumar Vishvakarma. An accurate and noninvasive skin cancer screening based on imaging technique. *International Journal of Imaging Systems and Technology*, 32(1):354–368, 2021. Funding information: Council of Scientific and Industrial Research (CSIR) New Delhi, Government of India, Grant/Award Number: 09/1022(0026)/2016-EMR-I.

[7]Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

[8]Philipp Tschandl, Noel Codella, Banu N. Akay, Giuseppe Argenziano, Ralph P. Braun, Horacio Cabo, David Gutman, Allan Halpern, Tomas Johnson, Harald Kittler, Aimilios Lallas, Caterina Longo, Josep Malvehy, Elvira Moscarella, Johan Paoli, Susana Puig, Cliff Rosendahl, Pietro Rubegni, Alon Scope, H. Peter Soyer, Luc Thomas, Iris Zalaudek, and Ashfaq A. Marghoob. Comparison of the accuracy

of human readers versus machine-learning algorithms for pigmented skin lesion classification. *The Lancet Oncology*, 19(5):e238–e246, 2018.

[9]Md Manjurul Ahsan, Stacey A Luna, and Zahed Siddique. Machine-learning-based disease diagnosis: A comprehensive review. *Healthcare*, 10(3):541, 2022.

[10]H.A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. Ben Hadj Hassen, L. Thomas, A. Enk, L. Uhlmann, Christina Alt, Monika Arenbergerova, Renato Bakos, Anne Baltzer, Ines Bertlich, Andreas Blum, Therezia Bokor-Billmann, Jonathan Bowling, Naira Braghiroli, Ralph Braun, Kristina Buder-Bakhaya, Timo Buhl, Horacio Cabo, Leo Cabrijan, Naciye Cevic, Anna Classen, David Deltgen, Christine Fink, Ivelina Georgieva, Lara-Elena Hakim-Meibodi, Susanne Hanner, Franziska Hartmann, Julia Hartmann, Georg Haus, Elti Hoxha, Raimonds Karls, Hiroshi Koga, Jürgen Kreusch, Aimilios Lallas, Pawel Majenka, Ash Marghoob, Cesare Massone, Lali Mekokishvili, Dominik Mestel, Volker Meyer, Anna Neuberger, Kari Nielsen, Margaret Oliviero, Riccardo Pampena, John Paoli, Erika Pawlik, Barbar Rao, Adriana Rendon, Teresa Russo, Ahmed Sadek, Kinga Samhaber, Roland Schneiderbauer, Anissa Schweizer, Ferdinand Toberer, Lukas Trennheuser, Lyobomira Vlahova, Alexander Wald, Julia Winkler, Priscila Wölbing, and Iris Zalaudek. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8):1836–1842, 2018. Immune-related pathologic response criteria.

[11]V. Adarsh, G. R. Gangadharan, Ugo Fiore, and Paolo Zanetti. Multimodal classification of alzheimer's disease and mild cognitive impairment using custom mkscddl kernel over cnn with transparent decision-making for explainable diagnosis. *Scientific Reports*, 14, 01 2024.

[12]Meiping Jiang, Sanlin Lei, Junhui Zhang, Liqiong Hou, Zhang Meixiang, and Yingchun Luo. Multimodal imaging of target detection algorithm under artificial intelligence in the diagnosis of early breast cancer. *Journal of Healthcare Engineering*, 2022:1–10, 01 2022.

[13]Santosh Kumar, Sachin Kumar Gupta, Vinit Kumar, Manoj Kumar, Mithilesh Kumar Chaube, and Nenavath Srinivas Naik. Ensemble multimodal deep learning for early diagnosis and accurate classification of covid-19. *Computers and Electrical Engineering*, 103:108396, 2022.

[14]Mehak Arshad, Muhammad Attique Khan, Usman Tariq, Ammar Armghan, Fayadh Alenezi, Muhammad Younus Javed, Shabnam Mohamed Aslam, and Seifedine Kadry. A computer-aided diagnosis system using deep learning for multiclass skin lesion classification. *Computational Intelligence and Neuroscience*, 2021:9619079, 2021. 15 pages.

[15]Onur Sevli. A deep convolutional neural network-based pigmented skin lesion classification application and experts evaluation. *Neural Computing and Applications*, 33(18):12039–12050, 2021.

[16]Gang Wang, Pu Yan, Qingwei Tang, Lijuan Yang, and Jie Chen. Multiscale feature fusion for skin lesion classification. *BioMed Research International*, 2023(1):5146543, 2023.

[17]MS AbdulmateenAdebiyi, MS NaderAbdalnabi, Emily Hoffman Smith, Jesse Hirner, Eduardo J. Simoes, Mirna Becevic, and Praveen Rao. Accurate skin lesion classification using multimodal learning on the ham10000 dataset. In *medRxiv*, 2024.

[18]Varun Srivastava, Deepika Kumar, and Sudipta Roy. A median based quadrilateral local quantized ternary pattern technique for the classification of dermatoscopic images of skin cancer. *Comput. Electr. Eng.*, 102(C), September 2022.

[19]Soumyya Kanti Datta, Mohammad Abuzar Shaikh, Sargur N. Srihari, and Mingchen Gao. Soft-attention improves skin cancer classification performance, 2021.

[20]Zhangli Lan, Songbai Cai, Xu He, and Xinpeng Wen. Fixcaps: An improved capsules network for diagnosis of skin cancer. *IEEE Access*, PP:1–1, 2022.

[21]Nils Gessert, Maximilian Nielsen, Mohsin Shaikh, René Werner, and Alexander Schlaefer. Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. *MethodsX*, 7:100864, 2020.

[22]Chubin Ou, Sitong Zhou, Ronghua Yang, Weili Jiang, Haoyang He, Wenjun Gan, Wentao Chen, Xinchi Qin, Wei Luo, Xiaobing Pi, and Jiehua Li. A deep learning based multimodal fusion model for skin lesion diagnosis using smartphone collected clinical images and metadata. *Frontiers in Surgery*, 9, 2022.

[23]David Restrepo, Chenwei Wu, Sebastián Andrés Cajas, Luis Filipe Nakayama, Leo Anthony Celi, and Diego M López. Multimodal deep learning for low-resource settings: A vector embedding alignment approach for healthcare applications. *medRxiv*, 2024.

[24]National Cancer Institute. Melanoma treatment (pdq®)–health professional version. National Cancer Institute, 2024. Accessed: 2025-01-09.

[25]William D. James, Timothy G. Berger, and Dirk M. Elston. *Andrews' Diseases of the Skin: Clinical Dermatology*. Saunders Elsevier, 10th edition, 2006.

[26]National Cancer Institute. Skin cancer treatment (pdq®)–health professional version. National Cancer Institute, 2024. Accessed: 2025-01-09.

[27]C. P. H. Reinehr and R. M. Bakos. Actinic keratoses: review of clinical, dermoscopic, and therapeutic aspects. *Anais Brasileiros de Dermatologia*, 94(6):637–657, 2019.

[28]R. Scott and A. Oakley. Benign keratosis: A useful term? *Dermatology Practical & Conceptual*, 13(2):e2023115, 2023. Advance online publication.

[29]D. J. Myers and E. P. Fillman. Dermatofibroma. 2024. Updated 2024 Feb 29. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2025 Jan-.

[30]J. E. Steiner and B. A. Drolet. Classification of vascular anomalies: An update. *Seminars in Interventional Radiology*, 34(3):225–232, 2017.

[31]Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998.

[32]Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012.

[33]Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[34]Jiazhi Liang. Image classification based on resnet. *Journal of Physics: Conference Series*, 1634(1):012110, 2020.

[35]Guoping Xu, Xiaxia Wang, Xinglong Wu, Xuesong Leng, and Yongchao Xu. Development of skip connection in deep neural networks for computer vision and medical image analysis: A survey. *arXiv preprint arXiv:2405.01725*, 2024.

[36]Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, pages 689–696. Omnipress, 2011.

[37]Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. Learning visually grounded sentence representations. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 408–418, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[38]Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1958–1966, 2016.

[39]Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[40]Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468, Austin, Texas, November 2016. Association for Computational Linguistics.

[41]John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. Gated multimodal networks. *Neural Comput. Appl.*, 32(14):10209–10228, July 2020.

[42]Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of NeurIPS 2017*, 2017.

[43]Chen Lu, Dhruv Batra, Devi Parikh, and Stephen Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of NeurIPS 2019*, 2019.

[44]J. Dinnes, J. J. Deeks, M. J. Grainge, N. Chuchu, L. Ferrante di Ruffano, R. N. Matin, D. R. Thomson, K. Y. Wong, R. B. Aldridge, R. Abbott, M. Fawzy, S. E. Bayliss, Y. Takwoingi, C. Davenport, K. Godfrey, F. M. Walter, H. C. Williams, and Cochrane Skin Cancer Diagnostic Test Accuracy Group. Visual inspection for diagnosing cutaneous melanoma in adults. *The Cochrane database of systematic reviews*, 12(12):CD013194, 2018.

[45]Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1):310, 2020.

[46]Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[47]Joanna Jaworek-Korjakowska, Andrzej Brodzicki, Bill Cassidy, Moi Hoon Yap, et al. Interpretability of a deep learning based approach for the classification of skin lesions into main anatomic body sites. *Cancers*, 13(23):6048, 2021.

[48]Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[49]Katherine Mader. Skin cancer mnist: Ham10000. Kaggle, 2018. Accessed: 2025-01-09.

FIGURE CAPTIONS

Table 1. Summary of Skin Lesion Classification Studies

Author	Year	Dataset	Classifiers/Preprocessors	Performance Achieved
Arshad et al.	2021	HAM100000	ResNet-50, ResNet-101, Ensemble Subspace Discriminative Classifiers	Achieved 95% accuracy. Relies on dermoscopic images and computationally expensive fusion methods.
Sevli et al.	2021	HAM10000	CNN, Image Resizing, Contrast Enhancement, Softmax Activation	Achieved 91.51% accuracy. Limited by non-multimodal approach and performance issues with dark skin tones and hair artifacts.
Wang et al.	2023	HAM10000	Two-Stream Network (DenseNet-121, VGG-16), Multireceptive Field Module, GeM	Achieved 91.24% accuracy. Non-lightweight architecture limits deployment in resource-constrained settings.
Adebiyi et al.	2024	Not specified	ALBEF Framework, Vision Transformer, BERT	Achieved 94.11% accuracy and AUC-ROC of 0.9426. Potential for improvement with additional clinical features.
Srivastava et al.	Not specified	HAM10000, ISIC UDA	M-QuadLTQP, Modified CNN	Achieved 96% accuracy. Computationally intensive and lacks multimodality.
Datta et al.	2021	HAM10000, ISIC 2017	Soft-Attention Mechanism, Inception ResNet v2, ResNet, DenseNet	Achieved 93.7% precision. Not multimodal and lacks generalizability testing.
Lan et al.	Not specified	HAM10000	FixCaps, Large-Kernel Convolution, CBAM	Achieved 96.49% accuracy. Not multimodal and underperformed on VASC lesions.
Gessert et al.	Not specified	ISIC 2019	Ensemble of EfficientNets, SENet, ResNeXt WSL, Patient Metadata	Improved performance by 1–2%. Suffered from overfitting and reduced performance on unknown class.
Ou et al.	2022	PAD-UPES-20	Multimodal Model, Smartphone Images, Metadata	Achieved 76.8% accuracy. Limited by small dataset and low accuracy.
Restrepo et al.	2024	BRSET, HAM10000, SatelliteBench	Vector Embeddings (DINO v2, LLAMA 2, CLIP)	High accuracy on BRSET but lower on HAM10000 due to domain-specific challenges.

Table 2. Breakdown of Classes in the HAM10000 Dataset

Class Name	Number of Images	Description
Melanoma (MEL)	1,113	Melanoma is a malignant tumor of melanin-producing melanocyte cells [24]
Melanocytic nevus (NV)	6,705	Benign moles of pigment-producing skin cells [25]
Basal cell carcinoma (BCC)	514	Slow-growing, locally destructive skin cancer derived from the basal cell layer of the epidermis [26]
Actinic keratosis / Bowen's disease (AKIEC)	327	Precancerous scaly lesions found on sun-damaged skin [27]
Benign keratosis (BKL)	1,099	Common benign skin lesions with sharply demarcated borders, homogenous brown pigmentation, and fine scaling that include Seborrheic keratosis (SK), lichen planus-like keratosis (LPLK), and solar lentigo (SL) [28]
Dermatofibroma (DF)	115	Benign skin nodules of soft tissue [29]
Vascular lesions (VASC)	142	Lesions involving blood vessels, such as angiomas [30]

Table 3. Clinical Features in the HAM10000 Dataset

Clinical Feature	Distribution	Description
Diagnosis (dx)	- Nevus (nv): 67% - Melanoma (mel): 11% - Other types: 22%	Medical diagnosis of the skin lesion
Diagnostic Method	- Histopathology: 53% - Follow-up: 37% - Other methods: 10%	Method used to confirm the diagnosis
Patient Age	- Range: 0-85 years - Divided into 10-year intervals	Age of the patient at the time of diagnosis
Gender	- Male: 54% - Female: 45% - Unspecified: 1%	Patient's gender identification
Anatomical Location	- Back: 22% - Lower extremity: 21% - Other locations: 57%	Body location where the lesion was found

Table 4. Hardware and Software Configuration

Component	Specification
GPU	NVIDIA GeForce RTX 4060
Processor	AMD Ryzen 7 7800X3D
Memory	16GB DDR4 RAM
Operating System	Linux Mint 21.1
CUDA	Enabled

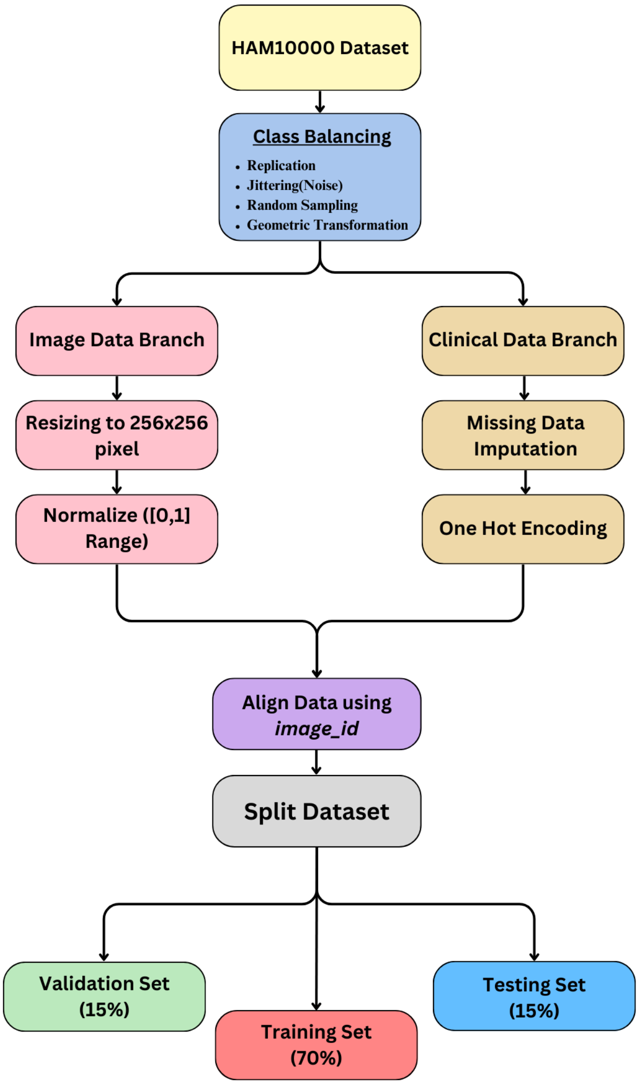

Figure 1. Overview of the preprocessing pipeline.

Table 5. Training Configuration

Parameter	Value
Batch Size	64
Number of Epochs	100
Learning Rate	0.001
Optimizer	Adam
Loss Function	Cross-Entropy Loss

Table 6. Performance of Individual Models in Ablation Studies

Model	Accuracy	Precision	Recall	F1-Score
Clinical CNN (Metadata Only)	77.0%	0.76	0.75	0.76
DermiResNet (Image Only)	92.0%	0.91	0.92	0.91

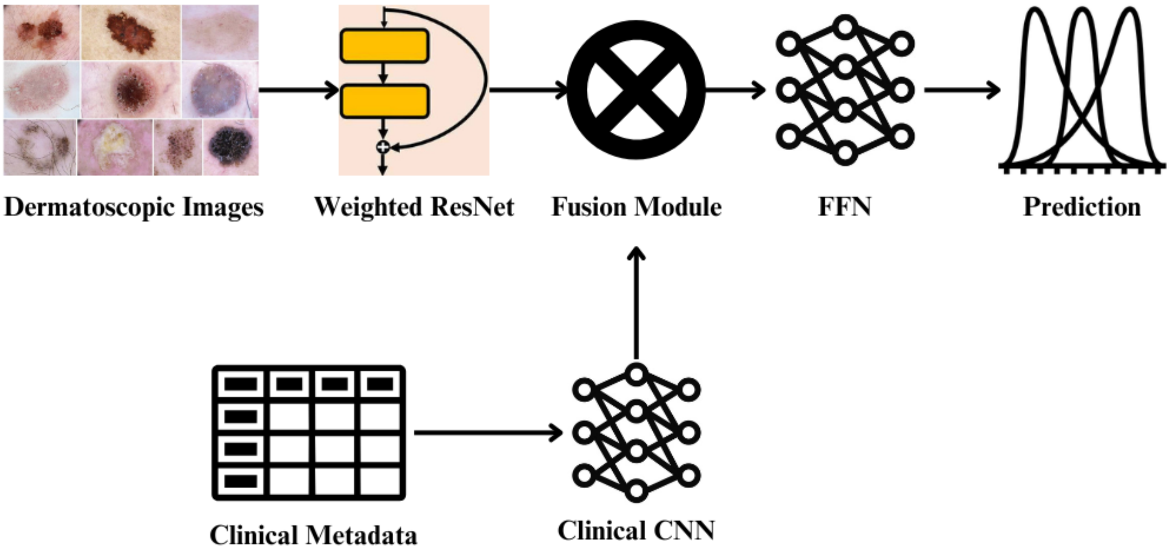

Figure 2. Overview of the multimodal fusion pipeline.

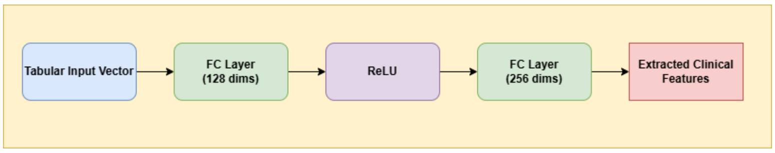

Figure 3. Architecture of the Clinical CNN

Table 7. Performance of Multimodal Fusion Techniques

Fusion Technique	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Simple Concatenation	96.5%	0.93	0.93	0.93
Weighted Concatenation	97.15%	0.97	0.97	0.97
Hadamard Product	98.85%	0.97	0.97	0.97
Tensor Fusion	96.52%	0.96	0.96	0.96
Bilinear Fusion	98.76%	0.98	0.98	0.98
Gated Fusion	93.0%	0.93	0.93	0.92
Self-Attention	92.70%	0.92	0.93	0.92
Cross-Attention	98.86%	0.98	0.98	0.98

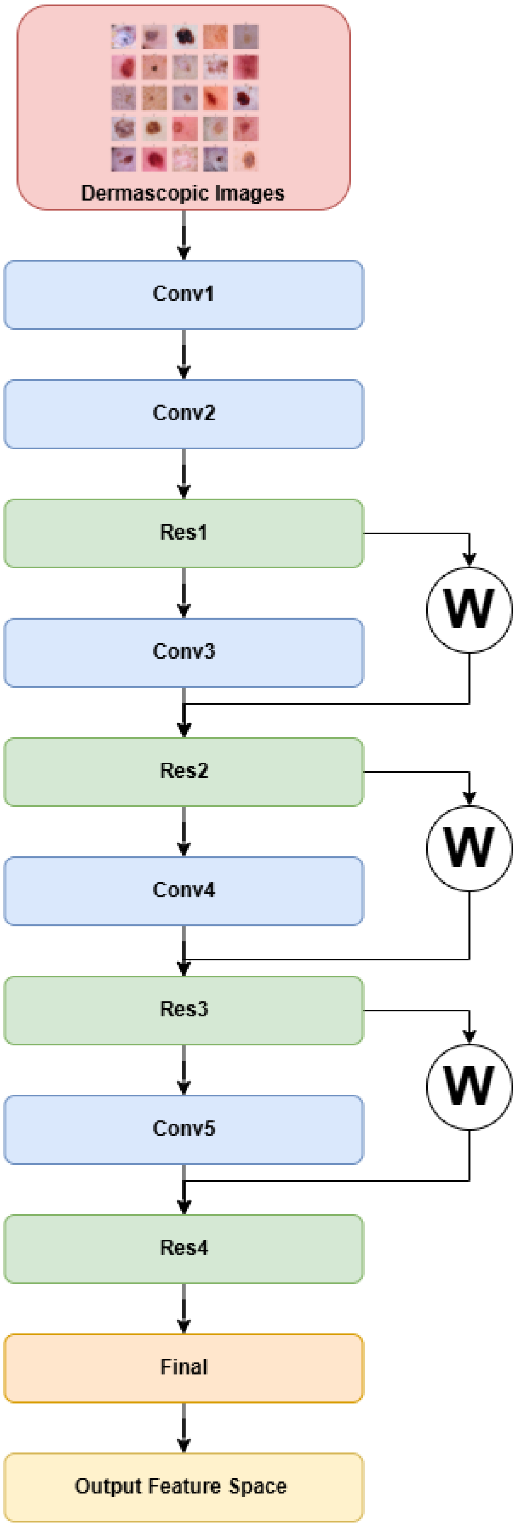

Figure 4. DermiResNet

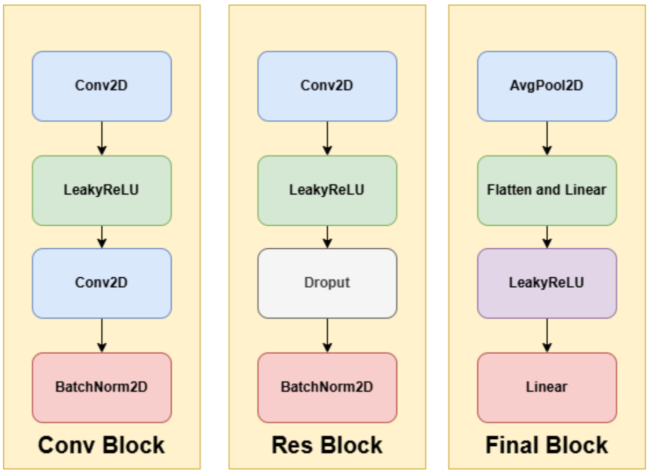

Figure 5. Blocks in DermiResNet

Table 8. Confusion Matrix of the Best Perfomring Cross-Attention Classifier

True Label	**bkl**	**bcc**	**df**	**mel**	**nv**	**vasc**	**akiec**	**Total**
Benign keratosis (bkl)	434	0	0	7	4	0	5	450
Basal cell carcinoma (bcc)	0	438	0	1	0	0	1	440
Dermatofibroma (df)	0	0	435	0	0	0	0	435
Melanoma (mel)	1	1	0	464	1	0	2	469
Melanocytic nevi (nv)	5	1	1	4	449	0	0	460
Vascular lesions (vasc)	0	0	0	0	0	434	0	434
Actinic keratoses (akiec)	0	0	0	0	0	0	460	460
Predicted Total	440	440	436	476	454	434	468	3148

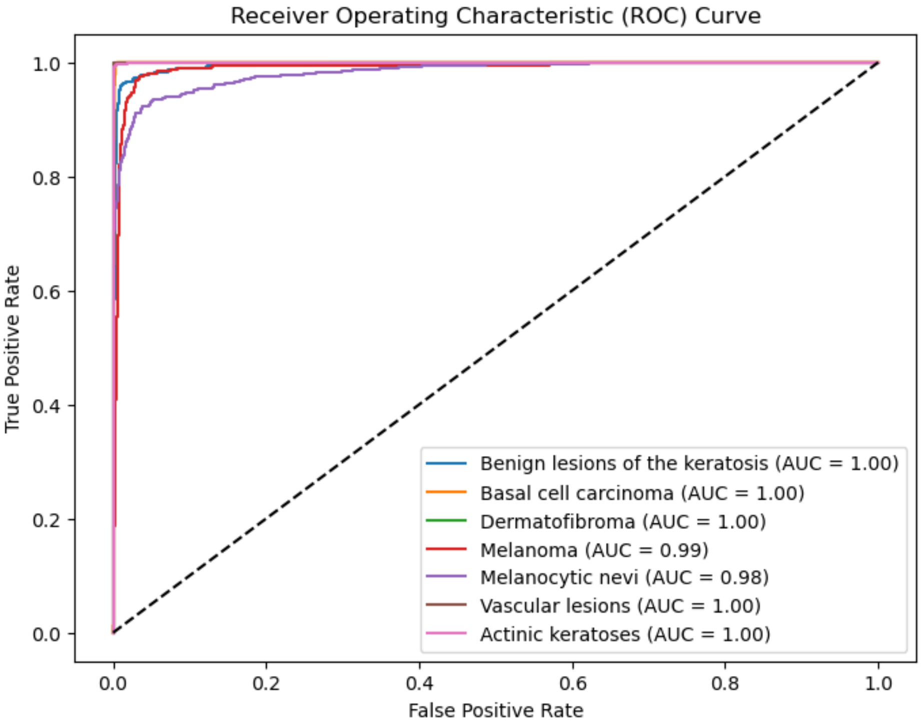

Figure 6. ROC-AUC curve for melanoma, nevi, and other classes.

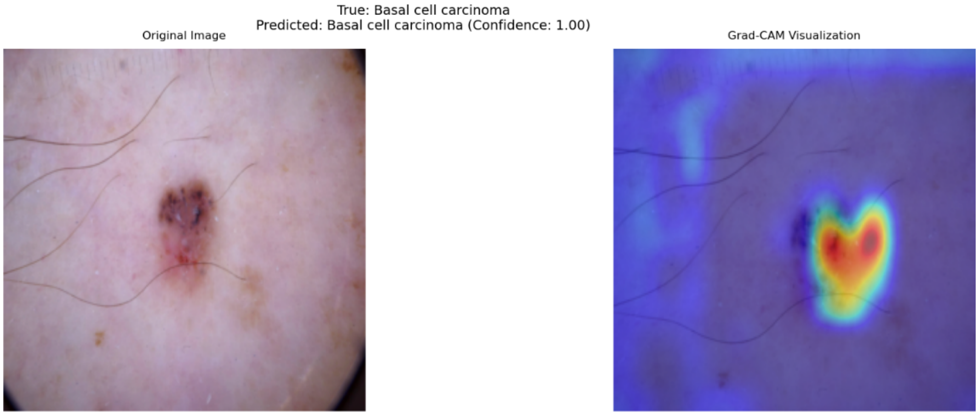
True: Basal cell carcinoma
Predicted: Basal cell carcinoma (Confidence: 1.00)
Original Image
Grad-CAM Visualization

Figure 7. Grad-CAM visualization for a dermatoscopic image classified as melanoma. The left panel shows the original image, while the right panel highlights clinically significant regions contributing to the model's prediction. Red regions indicate high importance, and blue regions indicate low importance.

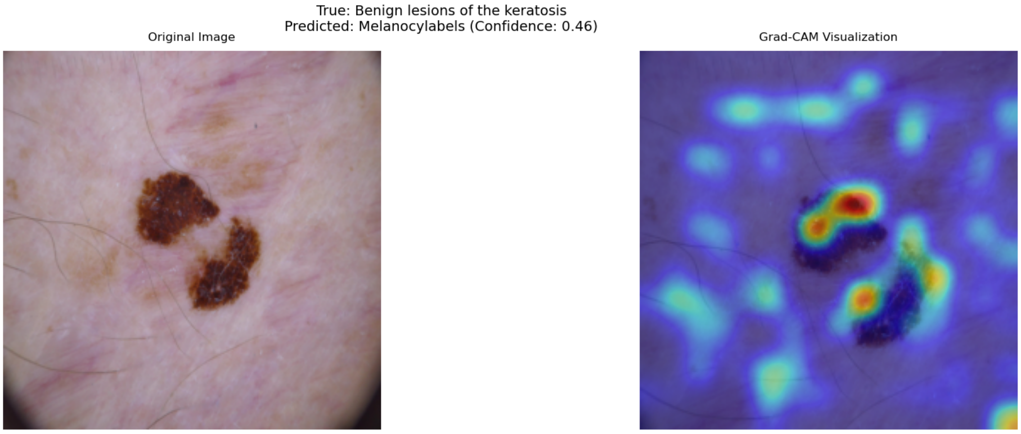
True: Benign lesions of the keratosis
Predicted: Melanocylabels (Confidence: 0.46)
Original Image
Grad-CAM Visualization

Figure 8. Grad-CAM visualization of a misclassified Benign Keratosis case. The model incorrectly predicts this lesion as Melanocytic (mel) with a confidence of 0.46.

Table 9. Instance-Specific Clinical Feature Importance for a Melanocytic Nevus Instance

Clinical Feature	Relative Importance
Dx Type: Follow-Up	1.00
Localization: Hand	0.80
Localization: Scalp	0.75
Localization: Neck	0.70
Localization: Acral	0.65
Localization: Lower Extremity	0.60
Localization: Chest	0.55
Localization: Unknown	0.50
Localization: Abdomen	0.45
Localization: Genital	0.40
Sex: Male	0.35
Dx Type: Consensus	0.30
Sex: Unknown	0.25
Localization: Upper Extremity	0.20
Dx Type: Confocal	0.15
Localization: Trunk	0.12
Localization: Face	0.10
Sex: Female	0.08
Localization: Back	0.06
Localization: Ear	0.04
Localization: Foot	0.03
Age	0.02
Dx Type: Histopathology	0.01

Figure 9. Other Grad-CAM Visualization with Cross Attention Fusion

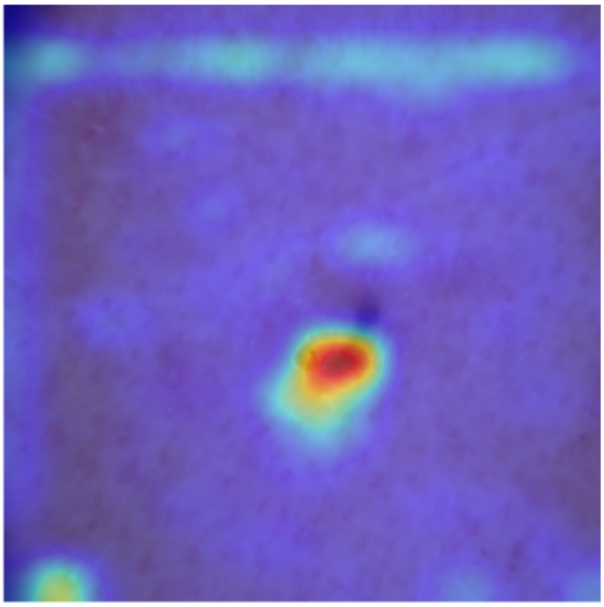

Figure 10. Basal Cell carcinoma

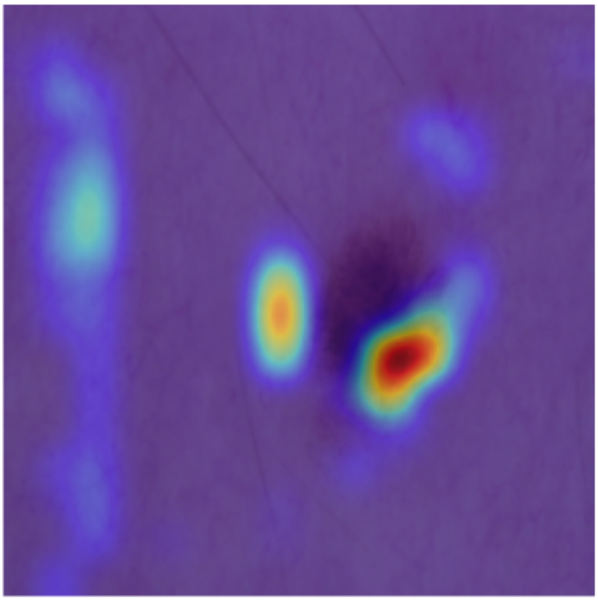

Figure 11. Melanocytic Nevi

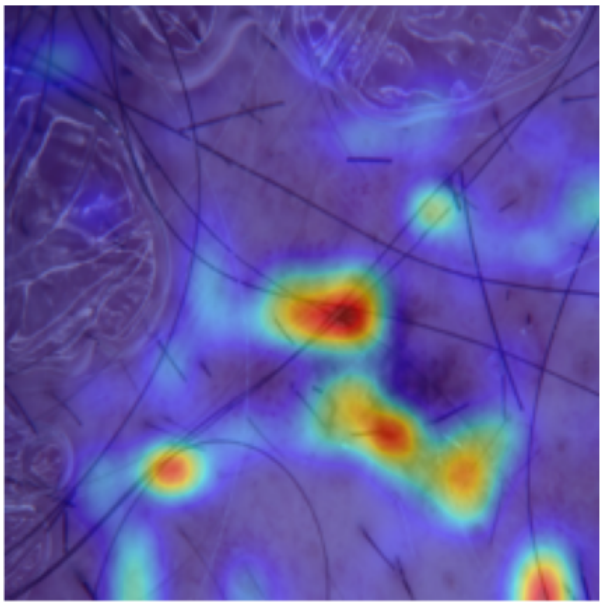

Figure 12. Vascular Lesions

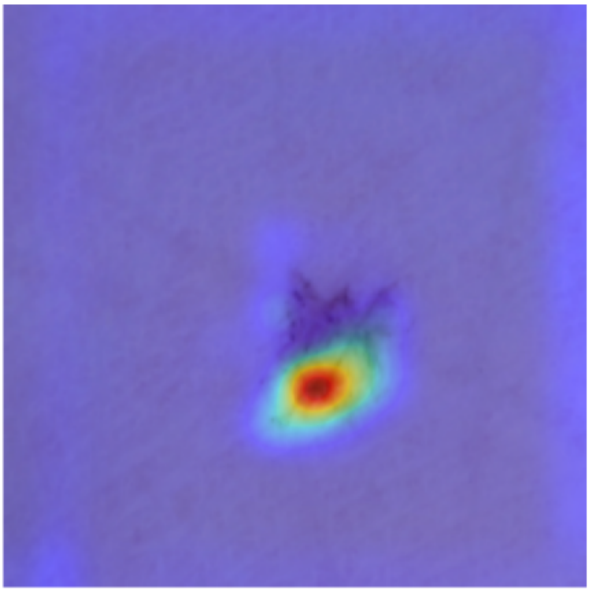

Figure 13. Melanoma