The Eastasouth Journal of Information System and Computer Science
Vol. 1, No. 02, December, pp. 279-299

Deep Learning for Accurate Classification of Seven Types of Skin Lesions from Dermoscopic Images

Sumayya Jahan[1], Kallol Chakraborty Shekhor[2], Md Sahid Hossain[3], Md Fakrul Alam[4], Abdur Rahim[5], Md Sazzad Hossain Shehan[6], Md Anwar Hossain[7], Md Abedur Rahman[8], Mohammed Adnan[9]

[1,7,8] Department of Computer Science, Maharishi International University, Fairfield, Iowa, USA
[2,4] Master of Science in Information Studies, Trine University, Allen Park, Michigan, USA
[3] Senior Software Engineer, Prime Tech Solutions Ltd., Dhaka, Bangladesh
[5,6] Bachelor of Science in Computer Science and Engineering, BRAC University, Dhaka, Bangladesh
[9] Department of Computer Science and Engineering, VIT-AP University, India

Article Info

Article history:

Received Dec, 2023
Revised Dec, 2023
Accepted Dec, 2023

Keywords:

ConvNeXt Tiny;
Dermoscopic imaging;
Gated feature fusion;
Pyramid pooling;
Skin lesion classification;
Swin Transformer

ABSTRACT

Accurate multiclass classification of dermatoscopic images remains challenging because skin lesions often exhibit substantial visual similarity, considerable variation in color and morphology, and severe class imbalance. This study presents a hybrid deep learning framework that integrates ConvNeXt Tiny, pyramid pooling context modeling, and a hierarchical Swin Transformer for seven class skin lesion classification. The ConvNeXt branch extracts progressive local texture, color, and morphological representations, while the pyramid pooling module aggregates lesion context at spatial scales defined by pooling bins of 1, 2, 3, and 6. In parallel, the Swin Transformer branch models hierarchical spatial relationships and long range dependencies through shifted window attention. The three branch representations are globally pooled, projected into a shared 512 dimensional latent space, and combined using a learned gated attention mechanism that adaptively controls their contributions to the final prediction. The framework was evaluated on the HAM10000 dataset containing 10,015 dermatoscopic images from seven diagnostic categories. Images were partitioned at the lesion level into 7,010 training, 1,502 validation, and 1,503 testing samples to prevent information leakage between images of the same lesion. Class balanced focal loss, AdamW optimization, ImageNet normalization, and controlled image augmentation were used during training. The proposed model achieved an accuracy of 95.81%, balanced accuracy of 92.51%, macro precision of 92.64%, macro recall of 92.51%, macro F1 score of 92.58%, and macro area under the receiver operating characteristic curve of 99.01%. Ablation analysis showed progressive improvements after introducing pyramid pooling, Swin features, channel attention, and gated fusion, with a macro F1 increase of 3.21 percentage points over the ConvNeXt Tiny baseline. These findings indicate that complementary local, contextual, and global representations can improve image level skin lesion classification, particularly under substantial class imbalance.

Corresponding Author:

Name: Mohammed Adnan
Institution: Department of Computer Science, Maharishi International University, Fairfield, Iowa, USA
Email: drmohammedadnan110@gmail.com

1. INTRODUCTION

Skin cancer represents a major global health concern, and its clinical management depends strongly on the timely recognition of suspicious lesions. Dermoscopy is a noninvasive imaging technique that improves visualization of subsurface structures, pigmentation patterns, vascular arrangements, and lesion boundaries that may not be fully observable through routine visual inspection. Despite its clinical value, interpretation of dermatoscopic images remains challenging because different lesion categories can share similar colors, textures, shapes, and structural patterns. In addition, lesions belonging to the same diagnostic category may exhibit considerable visual variation across patients, anatomical locations, acquisition devices, and imaging conditions. These factors can increase diagnostic uncertainty, particularly when distinguishing melanoma from benign melanocytic nevi or separating keratinocytic lesions with overlapping morphological characteristics [1,2]. Computer assisted analysis of dermatoscopic images has therefore become an active research area. Early approaches relied on manually designed descriptors for lesion color, border irregularity, texture, symmetry, and geometric structure. Although these methods provided interpretable representations, their effectiveness depended heavily on accurate lesion segmentation and careful feature engineering. Their ability to represent complex visual interactions was also limited. Deep convolutional neural networks substantially improved automated skin lesion analysis by learning hierarchical representations directly from image data. Architectures such as ResNet, DenseNet, and EfficientNet have demonstrated strong performance in multiclass dermatoscopic image classification by progressively encoding local edges, textures, colors, and higher level lesion structures [3,4]. More recent convolutional architectures have further improved representation quality. ConvNeXt modernizes conventional convolutional networks using design principles associated with contemporary vision models while preserving the spatial inductive biases of convolution. Its hierarchical feature extraction process is well suited to dermatoscopic images because local pigmentation, fine texture, vascular structures, and lesion borders often provide important diagnostic evidence. However, a conventional convolutional backbone may not fully represent contextual patterns distributed across distant image regions. Successive local operations can gradually increase the receptive field, but the resulting representation may remain dominated by local feature aggregation. This limitation is important in skin lesion classification, where the relationship between central lesion structures, peripheral pigmentation, asymmetry, and surrounding tissue may contribute to class discrimination.

Multiscale contextual modeling provides one possible approach to addressing this limitation. Skin lesions vary considerably in size, shape, and spatial distribution. A fixed receptive field may emphasize only one level of structure and may therefore overlook diagnostically relevant information appearing at different scales. Pyramid pooling addresses this problem by aggregating features from multiple spatial partitions. Pooling bins representing global and regional image contexts can encode both overall lesion organization and localized structural patterns. When combined with deep convolutional features, pyramid pooling can strengthen scale awareness without requiring multiple independently processed image resolutions [5]. Nevertheless, pooled contextual representations alone may not sufficiently capture flexible dependencies between distant lesion regions. Vision transformers provide a complementary strategy by using attention based interactions to model relationships between spatial tokens. The Swin Transformer introduces hierarchical feature learning and shifted window attention, allowing the model to capture both local interactions and progressively broader spatial dependencies

with manageable computational complexity [6]. This design is relevant to dermatoscopic analysis because lesion interpretation often depends on relationships among spatially separated structures, including irregular pigmentation, asymmetric regions, peripheral streaks, and distributed vascular patterns. However, transformer representations can be data intensive and may not preserve local texture biases as effectively as convolutional models when training data are limited or severely imbalanced. Consequently, exclusive dependence on either convolutional or transformer based representations may not provide the most effective balance between local detail and global context.

Several studies have combined convolutional networks and transformers for medical image classification. In many cases, features are merged using direct concatenation or elementwise addition. These operations assume that every representation contributes equally or that a subsequent dense layer can implicitly learn the relative importance of each branch. Such fusion strategies may be suboptimal when the branches encode heterogeneous information. For example, ConvNeXt features emphasize local texture and morphology, pyramid pooling features represent multiscale context, and Swin features model hierarchical spatial dependencies. The usefulness of these representations may vary across lesion categories and individual images. A melanoma image with irregular global asymmetry may benefit more from transformer based context, whereas a vascular lesion may depend more strongly on fine local structures. An adaptive fusion mechanism is therefore needed to control branch contributions according to the information contained in each image. The class distribution of commonly used dermatoscopic datasets presents an additional challenge. The HAM10000 dataset contains 10,015 images across seven diagnostic categories, but melanocytic nevi account for most samples, while dermatofibroma and vascular lesions are represented by relatively few images. A model optimized primarily for overall accuracy may consequently favor the majority class and provide inadequate recognition of clinically important minority categories. Evaluation based only on accuracy can also conceal substantial differences in classwise sensitivity. Balanced accuracy, macro precision, macro recall, macro F1 score, and classwise area under the receiver operating characteristic curve are therefore required for a more informative assessment. Training strategies should also account for imbalance through class aware loss functions and controlled augmentation.

Another methodological concern involves partitioning images from the same physical lesion. HAM10000 can contain multiple images associated with one lesion identifier. Random image level splitting may place related images in both training and testing subsets, allowing the model to encounter highly similar lesion characteristics during optimization and evaluation. This overlap can produce optimistic performance estimates. Lesion level partitioning is therefore necessary to ensure that all images associated with one lesion remain within a single subset. Such separation provides a more reliable assessment of image level classification performance on previously unseen lesions.Motivated by these considerations, this study proposes a hybrid framework that integrates ConvNeXt Tiny, pyramid pooling context modeling, and a four stage Swin Transformer for seven class dermatoscopic image classification. The ConvNeXt branch extracts local texture, color, and morphological representations through hierarchical convolutional stages. A pyramid pooling context module operates on the deep convolutional feature map using pooling bins of 1, 2, 3, and 6 to encode global and regional lesion context. In parallel, the Swin Transformer branch captures hierarchical spatial relationships through shifted window attention. The three feature streams are globally pooled and projected into a common 512 dimensional latent space. A learned gated attention mechanism then estimates the relative contribution of each branch and

produces an adaptively fused representation for final classification.

The principal contributions of this study are fourfold. First, a complementary three stream architecture is developed to jointly represent local dermatoscopic texture, multiscale lesion context, and hierarchical global dependencies. Second, a gated fusion mechanism is introduced to adaptively combine heterogeneous branch features rather than relying only on direct concatenation. Third, the model is evaluated using lesion level training, validation, and testing partitions to reduce information leakage between related images. Fourth, comprehensive evaluation is conducted using overall, balanced, macro averaged, classwise, and ablation based analyses to assess performance under severe class imbalance. The proposed framework achieved 95.81 percent accuracy, 92.51 percent balanced accuracy, 92.58 percent macro F1 score, and 99.01 percent macro area under the receiver operating characteristic curve on the held out test set. These results suggest that adaptive integration of local, contextual, and global representations can provide a useful approach to multiclass dermatoscopic image classification.

2. RELATED WORK

Machine learning and deep learning have been widely investigated for automated skin lesion analysis. Das et al. reviewed the application of machine learning in dermatology and emphasized its potential to support early skin cancer recognition when access to specialists is limited [1]. Similar artificial intelligence methods have been applied to cervical cancer analysis, demonstrating the broader relevance of computational models in cancer screening and classification [2]. Dildar et al. reviewed deep learning techniques for skin cancer detection and identified lesion color, symmetry, texture, shape, and border characteristics as important visual indicators [3]. Artificial intelligence based predictive analysis has also been studied for early chronic disease detection and personalized medicine [4]. Reviews by Goceri [5] and Nahata and Singh [6] further highlighted the importance of reliable image representation, transfer learning, and early lesion recognition. Recent medical image classification studies have increasingly used attention, multiscale representation, and lightweight architectures. Shekhor et al. combined multiscale attention fusion with a clinical recommendation assistant for brain tumour magnetic resonance image classification [7]. Manik investigated predictive multiomics analysis for Parkinson disease related neurosurgical applications [8], while Hossain et al. examined lightweight artificial intelligence models intended for limited hospital hardware [9]. Machine learning based biomarker analysis has also been explored for early ischemic stroke detection [10]. These studies indicate that complementary representations and computationally practical architectures can strengthen medical image and clinical data analysis.

Traditional skin lesion classification methods commonly relied on preprocessing, segmentation, and handcrafted feature extraction. Monika et al. combined hair removal, filtering, color clustering, ABCD characteristics, and texture features with a multiclass support vector machine [11]. Related precision oncology research has shown that integrating heterogeneous biomedical information may improve disease specific prediction and treatment planning [12]. Vidya and Karki used geodesic active contour segmentation together with ABCD, histogram of oriented gradients, and texture features for benign and melanoma classification [13]. Zghal and Derbel classified lesions using image enhancement, segmentation, and weighted ABCD measurements [14]. Kumar et al. combined fuzzy c means segmentation, color and texture features, and a differential evolution optimized artificial neural network [15]. Although these methods produced encouraging results, their performance depended strongly on segmentation quality and manually selected descriptors. Deep convolutional networks reduced the

dependence on handcrafted features by learning representations directly from dermatoscopic images. Gouda et al. evaluated a custom convolutional network and several transfer learning models after image enhancement, augmentation, normalization, and resizing [16]. Deep learning has also been applied to real time cardiovascular monitoring using wearable health information, demonstrating its usefulness across different medical data types [17]. Ameri used AlexNet based transfer learning for melanoma and nonmelanoma classification and reported an area under the curve of 0.91 [18]. Thurnhofer Hemsi and Domínguez compared several convolutional networks for seven class HAM10000 classification and found that DenseNet201 provided strong performance [19]. Large scale predictive analysis has additionally been studied for antibiotic resistance surveillance [20]. Rajarajeswari et al. applied transfer learning and end to end convolutional classification to dermatoscopic images while considering preprocessing and mobile deployment [21].

More recent studies have investigated cross dataset evaluation and model combination. Reis et al. proposed InSiNet and evaluated it using ISIC 2018, ISIC 2019, and ISIC 2020, reporting accuracies of 94.59%, 91.89%, and 90.54%, respectively [22]. Generative artificial intelligence and large scale analytics have also been examined as tools for accelerating biomedical and pharmaceutical research [23]. Imran et al. combined VGG, CapsNet, and ResNet learners and showed that ensemble decisions could improve several classification measures compared with individual networks [24]. Strategic biotechnology based artificial intelligence models have similarly been discussed as mechanisms for improving biomedical innovation [25]. Pacheco and Krohling demonstrated that combining skin images with patient clinical information improved balanced accuracy by approximately 7%, indicating that complementary information sources may strengthen automated skin cancer analysis [26]. Despite these developments, important limitations remain. Handcrafted approaches depend on accurate preprocessing and segmentation, while individual convolutional networks may not fully represent relationships among distant lesion regions. Transformer models can provide broader spatial modeling, but they may be affected by limited and imbalanced medical datasets. Conventional ensembles also increase computational complexity and often combine predictions without learning the relative importance of different feature sources. The present study therefore integrates ConvNeXt based local features, pyramid based multiscale context, and Swin Transformer representations within a gated fusion framework. This design enables adaptive combination of local, contextual, and global lesion information while maintaining lesion level data separation and balanced multiclass evaluation.

3. METHODOLOGY OVERVIEW

The proposed study was designed as an image level multiclass classification framework for recognizing seven categories of skin lesions from dermatoscopic images. The complete methodology consisted of dataset preparation, lesion level partitioning, image preprocessing, multibranch feature extraction, adaptive feature fusion, model optimization, and performance evaluation. The HAM10000 dataset was used as the primary experimental dataset. It contains 10,015 RGB dermatoscopic images representing actinic keratosis and intraepithelial carcinoma, basal cell carcinoma, benign keratosis like lesions, dermatofibroma, melanoma, melanocytic nevus, and vascular lesions. Because multiple images may correspond to the same physical lesion, the lesion identifier was used as the grouping variable during data partitioning. All images associated with one lesion were assigned to only one subset. The resulting dataset contained 7,010 training images, 1,502 validation images, and 1,503 testing images. This lesion level separation reduced the

possibility of information leakage between the training and evaluation subsets.

Each image was converted to RGB format and resized to a spatial resolution of 224×224 pixels using aspect preserving resizing followed by symmetric padding. ImageNet mean and standard deviation values were used for normalization. During training, controlled augmentation was applied using horizontal and vertical flipping, rotation, scaling, translation, color modification, and mild Gaussian blurring. No augmentation was applied to the validation or testing images. The proposed model contained three complementary feature streams. The first stream used a ConvNeXt Tiny encoder to extract hierarchical local representations related to lesion color, texture, borders, and morphology. Its four stages produced feature maps with channel dimensions of 96, 192, 384, and 768. The second stream used a pyramid pooling context module with pooling bins of 1, 2, 3, and 6. This module processed deep convolutional features to capture lesion information at multiple spatial scales and generated a 7×7×256 contextual representation. The third stream used a four stage Swin Transformer to model hierarchical spatial relationships and long range dependencies through shifted window attention.

Global average pooling converted the final ConvNeXt, pyramid context, and Swin feature maps into compact feature vectors. Since the three branches produced representations with different dimensions, independent projection layers mapped them into a common 512 dimensional latent space. A gated attention fusion module then estimated the relative contribution of each branch and combined the projected features into a unified representation. The fused feature was passed through normalization, a dense layer, activation, dropout, and a seven unit classification layer. Softmax probabilities were calculated during inference to determine the predicted diagnostic category. The model was trained using class balanced focal loss to reduce the influence of severe class imbalance. AdamW optimization was used with separate learning rates for the pretrained backbones and the fusion head. Training was performed for a maximum of 100 epochs using cosine learning rate annealing, a five epoch warm up period, gradient clipping, and early stopping based on validation macro F1. Experiments were repeated using five random seeds. Performance was evaluated using accuracy, balanced accuracy, macro precision, macro recall, macro F1 score, classwise sensitivity, specificity, and area under the receiver operating characteristic curve. Ablation experiments were also conducted to determine the contribution of pyramid pooling, Swin features, channel attention, direct concatenation, and gated fusion.

3.1 Dataset Description

The HAM10000 dataset was used to evaluate the proposed multiclass skin lesion classification framework. It contains 10,015 dermatoscopic RGB images collected from multiple populations and imaging systems and includes seven diagnostic categories. The reference labels were established using histopathological examination, confocal microscopy, expert consensus, or clinical follow up, depending on the lesion type. Each image was prepared for model development at an input resolution of 224×224×3. A major characteristic of the dataset is its substantial class imbalance, particularly for dermatofibroma and vascular lesions. The dataset also provides a lesion identifier, which was used as the grouping variable to ensure that images belonging to the same physical lesion remained within a single data partition. The principal characteristics of the dataset are summarized in Table 1. The class distribution is highly imbalanced, with melanocytic nevi representing 6,705 images, corresponding to 66.95 percent of the complete dataset. Melanoma and benign keratosis like lesions contain 1,113 and 1,099 images, respectively, whereas dermatofibroma and vascular lesions contain only 115 and 142 images. The

complete class distribution is presented in Table 2. To reduce the risk of information leakage, the dataset was divided at the lesion level into training, validation, and testing subsets. The resulting partitions contained 7,010 training images, 1,502 validation images, and 1,503 testing images, corresponding to approximately 70 percent, 15 percent, and 15 percent of the dataset, respectively. The class specific composition of each partition is reported in Table 3.

Table 1. HAM10000 dataset characteristics

Characteristic	Description
Dataset	Human Against Machine with 10000 training images
Abbreviation	HAM10000
Total images	10,015
Image modality	Dermatoscopic RGB images
Diagnostic categories	7
Primary task	Multiclass skin lesion classification
Image sources	Multiple populations and acquisition systems
Ground truth	Histopathology, confocal microscopy, expert consensus, or follow up
Recommended input size	224 x 224 x 3
Main challenge	Severe class imbalance, especially for DF and VASC
Grouping variable	Lesion identifier
Evaluation requirement	Images from the same lesion should remain in the same partition

Table 2. Class distribution of the HAM10000 dataset

Code	Diagnostic category	Images	Proportion
AKIEC	Actinic keratosis and intraepithelial carcinoma	327	3.27%
BCC	Basal cell carcinoma	514	5.13%
BKL	Benign keratosis like lesion	1,099	10.97%
DF	Dermatofibroma	115	1.15%
MEL	Melanoma	1,113	11.11%
NV	Melanocytic nevus	6,705	66.95%
VASC	Vascular lesion	142	1.42%
Total	All diagnostic categories	10,015	100.00%

Table 3. Lesion level dataset partition

Class	Training	Validation	Testing	Total
AKIEC	229	49	49	327
BCC	360	77	77	514
BKL	769	165	165	1,099
DF	80	17	18	115
MEL	779	167	167	1,113
NV	4,694	1,006	1,005	6,705
VASC	99	21	22	142
Total	7,010	1,502	1,503	10,015
Proportion	70.00%	15.00%	15.01%	100.00%

3.2 Dataset Preprocessing

All dermatoscopic images were converted to RGB format and prepared at a fixed spatial resolution of 224×224 pixels. To reduce geometric distortion, each image was resized while preserving its original aspect ratio, after which symmetric padding was applied to obtain the required square input dimension. Pixel values were normalized using the ImageNet channel means and standard deviations to maintain compatibility with the pretrained ConvNeXt Tiny and Swin Transformer backbones. Data

augmentation was applied only to the training subset and included horizontal and vertical flipping, rotation, scaling, translation, color jitter, and mild Gaussian blurring. These transformations were used to increase appearance variability and reduce overfitting while preserving diagnostically important lesion structures. No augmentation was applied to the validation or testing subsets to ensure consistent and unbiased performance evaluation. The complete preprocessing procedure is summarized in Table 4.Severe class imbalance was addressed using class balanced focal loss, which assigns greater importance to underrepresented categories and difficult samples. The model was optimized using AdamW with a learning rate of 3×10^{-5} for the pretrained backbone components and $3x10^{-4}$ for the fusion and classification layers. A weight decay of 0.05, batch size of 32, gradient clipping threshold of 1.0, and dropout rate of 0.30 were used during training. The learning rate was controlled through cosine annealing after a five epoch warm up period. Training continued for a maximum of 100 epochs, with early stopping applied when the validation macro F1 score did not improve for 15 consecutive epochs. The model checkpoint with the highest validation macro F1 score was selected for testing. To examine the stability of the results, experiments were repeated using random seeds 0, 1, 2, 3, and 4, as reported in Table 4.

Table 4. Preprocessing and training configuration

Component	Configuration
Input resolution	224 x 224 pixels
Image channels	RGB
Resizing	Aspect preserving resize followed by symmetric padding
Normalization	ImageNet mean and standard deviation
Training augmentation	Horizontal flip, vertical flip, rotation, scaling, translation, color jitter, and mild Gaussian blur
Validation augmentation	None
Test augmentation	None
Imbalance handling	Class balanced focal loss
Optimizer	AdamW
Backbone learning rate	3 x 10^-5
Fusion head learning rate	3 x 10^-4
Weight decay	0.05
Batch size	32
Maximum epochs	100
Learning rate scheduler	Cosine annealing
Warm up period	5 epochs
Early stopping patience	15 epochs
Dropout	0.30
Gradient clipping	1.0
Model selection criterion	Highest validation macro F1
Random seeds	0, 1, 2, 3, and 4

3.3 Proposed Model Architecture

The proposed framework integrates convolutional feature extraction, multiscale contextual modeling, transformer-based representation learning, and adaptive feature fusion for seven class skin lesion classification. A preprocessed dermatoscopic image of 224×224×3 pixels is simultaneously analyzed through complementary local, contextual, and global representation pathways. The complete architecture and the output

dimensions of its principal components are summarized in Table 5.

a. ConvNeXt Based Local Feature Extraction

The first branch employs a pretrained ConvNeXt Tiny encoder to extract hierarchical local features from the dermatoscopic image. ConvNeXt was selected because its modern convolutional design provides strong spatial inductive bias while supporting deep and computationally efficient feature learning. This branch primarily focuses on diagnostically relevant local characteristics, including pigmentation patterns, color variation, border irregularity, vascular structures, lesion texture, and morphological details. The encoder consists of four hierarchical stages. The first stage produces a feature map of 56×56×96, which preserves fine spatial information and represents low level texture, edge, and color patterns. The second stage reduces the spatial resolution to 28×28 while increasing the channel depth to 192, allowing the model to learn intermediate lesion structures. The third stage generates a 14×14×384 representation that captures more complex morphological patterns, including irregular lesion regions and heterogeneous pigmentation. The fourth stage produces a deep semantic feature map of 7×7×768, which summarizes high level lesion characteristics while retaining sufficient spatial organization for subsequent contextual analysis. The final ConvNeXt representation is used in two ways. First, global average pooling converts the deep feature map into a 768 dimensional local feature vector for adaptive fusion. Second, the spatial feature map is provided to the pyramid pooling context module, enabling the contextual branch to analyze the same deep convolutional representation at multiple spatial scales. The hierarchical outputs and corresponding functions of the ConvNeXt branch are presented in Table 5.

b. Pyramid Context and Swin Transformer Representation

The second component consists of the pyramid pooling context branch and the Swin Transformer branch. These pathways complement the local ConvNeXt representation by modeling multiscale lesion context and hierarchical spatial relationships. The pyramid pooling context module processes the deep ConvNeXt feature map using pooling bins of 1, 2, 3, and 6. The smallest pooling bin summarizes the complete lesion region and provides global contextual information, while the larger bin configurations preserve progressively finer regional details. These pooled representations allow the model to examine lesion appearance at multiple spatial scales, which is important because dermatoscopic abnormalities may occur globally across the lesion or within small localized regions. The pooled features are resized to a common spatial resolution, concatenated, and transformed through convolutional projection. Channel attention is then applied to emphasize informative contextual channels and suppress less relevant responses. The resulting context representation has an output size of 7×7×256, as reported in Table 5.

In parallel, the Swin Transformer branch captures hierarchical spatial dependencies through window based self attention and shifted window interactions. The input image is divided into nonoverlapping patches and embedded into a structured feature

representation. Four transformer stages then progressively reduce the spatial resolution while increasing the feature dimension. The first stage generates a 56×56×96 feature map and models relationships within local image windows. The second stage produces a 28×28×192 representation that captures broader contextual interactions. The third stage generates a 14×14×384 feature map and models dependencies between spatially separated lesion regions. The fourth stage produces a 7×7×768 semantic representation containing high level global information. Shifted window attention enables communication across neighboring windows while maintaining lower computational complexity than unrestricted global attention. This property allows the model to capture relationships among lesion centers, boundaries, peripheral structures, and surrounding skin regions. Global average pooling is subsequently applied to the final Swin feature map to obtain a 768 dimensional global feature vector. The contextual and transformer representations therefore provide information that is complementary to the local convolutional features.

c. Adaptive Feature Fusion and Classification

The final stage combines the local ConvNeXt vector, the pyramid context vector, and the global Swin Transformer vector. Global average pooling converts the three spatial representations into vectors with dimensions of 768, 256, and 768, respectively. Since these vectors differ in dimensionality and represent distinct types of information, separate projection layers map them into a common 512 dimensional latent feature space before fusion.

The projected features are initially concatenated to preserve information from all three branches. A gated attention mechanism then learns the relative importance of the local, contextual, and global representations. This mechanism allows the contribution of each branch to vary according to the visual characteristics of an individual lesion. Images dominated by fine texture or vascular patterns may depend more strongly on ConvNeXt features, whereas lesions with irregular global organization or distributed pigmentation may benefit more from the Swin representation. Similarly, lesions containing diagnostic structures at several spatial scales can receive a stronger contribution from the pyramid context branch.The adaptively combined representation is normalized using LayerNorm and transformed into a 512 dimensional fused feature vector. This vector is passed through a dense layer that reduces the feature dimension from 512 to 256, followed by a rectified linear activation and dropout with a probability of 0.30. The final dense layer maps the representation to seven output logits, corresponding to AKIEC, BCC, BKL, DF, MEL, NV, and VASC. Softmax is applied during inference to convert the logits into class probabilities, and the class with the highest probability is selected as the predicted diagnostic category. The pooling, fusion, and classification components are summarized in Table 5.

Table 5. Proposed ConvNeXt, pyramid pooling, and Swin fusion architecture

Stage	Module	Output representation	Primary function
Input	Preprocessed dermoscopic image	224 x 224 x 3	Provides a normalized lesion image

Stage	Module	Output representation	Primary function
Local Stage 1	ConvNeXt Tiny Stage 1	56 x 56 x 96	Extracts fine texture and color patterns
Local Stage 2	ConvNeXt Tiny Stage 2	28 x 28 x 192	Learns intermediate lesion structures
Local Stage 3	ConvNeXt Tiny Stage 3	14 x 14 x 384	Captures higher level morphological patterns
Local Stage 4	ConvNeXt Tiny Stage 4	7 x 7 x 768	Produces deep semantic features
Context Stage	Pyramid pooling with bins 1, 2, 3, and 6	7 x 7 x 256	Aggregates multiscale lesion context
Global Stage 1	Swin Transformer Stage 1	56 x 56 x 96	Models local window relationships
Global Stage 2	Swin Transformer Stage 2	28 x 28 x 192	Learns hierarchical global context
Global Stage 3	Swin Transformer Stage 3	14 x 14 x 384	Models distant lesion dependencies
Global Stage 4	Swin Transformer Stage 4	7 x 7 x 768	Produces global semantic features
Pooling	Global average pooling	768 + 256 + 768	Converts spatial feature maps into vectors
Fusion	Concatenation and gated attention	512	Selects complementary branch information
Classification	Dense layer, dropout, and Softmax	7 probabilities	Predicts the seven diagnostic classes

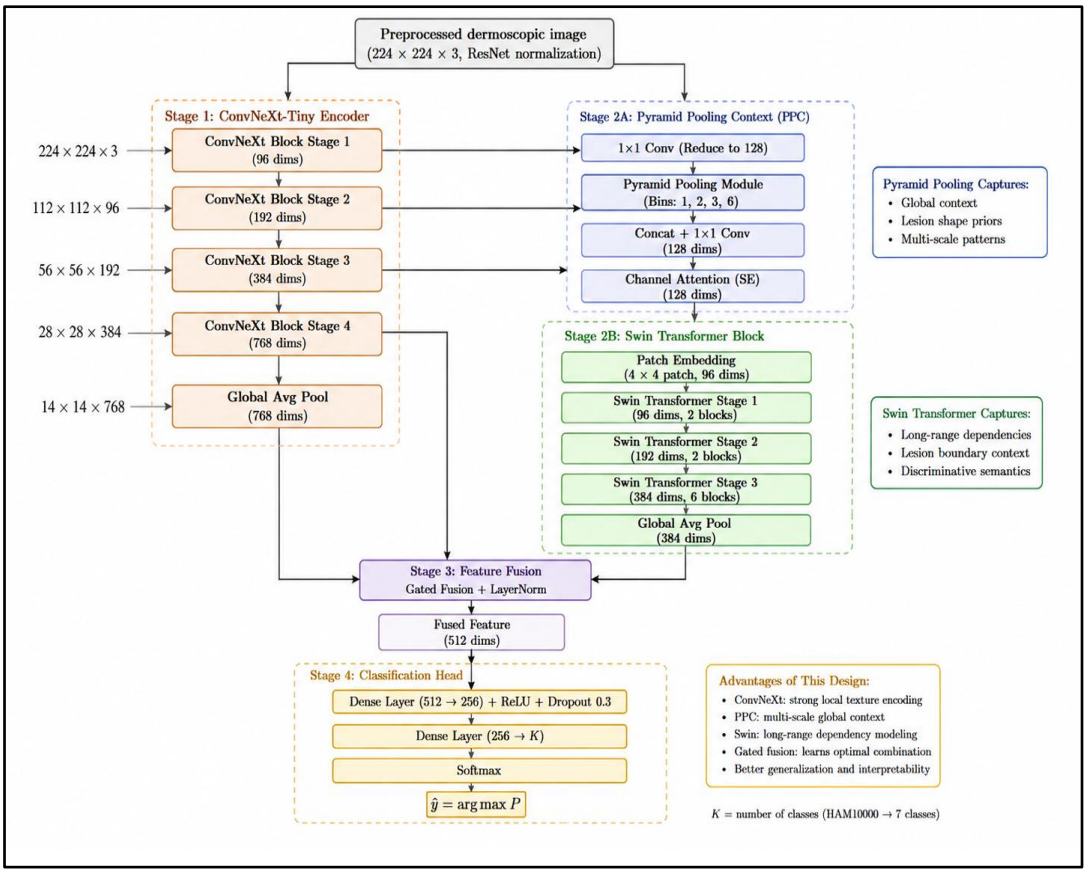

Figure 1. Architecture of the proposed ConvNeXt, pyramid pooling, and Swin Transformer fusion model for multiclass skin lesion classification

3.4 Baseline Models

Six representative deep learning models were selected as baselines to evaluate the effectiveness of the proposed architecture. ResNet50 was included as a conventional residual convolutional network that uses skip connections to support stable optimization and hierarchical feature learning. DenseNet121 was selected because its densely connected layers promote feature reuse and improve information flow across the network. EfficientNet B0 was used as a computationally efficient convolutional baseline that balances network depth, width, and input resolution through compound scaling. Together, these models provide established convolutional reference points for assessing the classification performance of the proposed framework.

Swin Tiny was included as a transformer based baseline to evaluate the contribution of hierarchical shifted window attention without convolutional feature fusion. It progressively models local and broader spatial relationships through window based self attention and patch merging. ConvNeXt Tiny was selected as the primary convolutional backbone baseline because it extracts strong local texture, color, boundary, and morphological representations using a modernized convolutional design. Comparing Swin Tiny and ConvNeXt Tiny separately also helps determine whether local convolutional features or hierarchical attention based features are more effective for HAM10000 skin lesion classification.An additional ConvNeXt Tiny with pyramid pooling baseline was developed to examine the specific contribution of multiscale contextual modeling. In this configuration, the final ConvNeXt feature map was processed using pyramid pooling bins of 1, 2, 3, and 6 before classification. This baseline captures both deep local features and contextual information at multiple spatial scales but does not include the Swin Transformer branch or gated fusion mechanism. All baseline models were trained and evaluated using the same lesion level partitions, preprocessing procedure, augmentation strategy, class imbalance treatment, optimization settings, model selection criterion, and random seeds. This consistent experimental protocol was used to ensure that the observed performance differences primarily reflected architectural variations rather than differences in data preparation or training conditions.

3.5 Hyperparameter Configuration

All baseline and proposed models were trained under a consistent hyperparameter configuration to support a fair architectural comparison. The pretrained feature extraction components were initialized using ImageNet weights and optimized with AdamW using a learning rate of $3\times\times10^{-5}$, while newly initialized classification, context, and fusion layers used a learning rate of $3\times\times10^{-4}$. For baseline models without context or fusion modules, the higher learning rate was applied only to the newly added classification head. A weight decay of 0.05, batch size of 32, dropout probability of 0.30, and gradient clipping threshold of 1.0 were used for all experiments. Class balanced focal loss was employed to reduce majority class dominance and improve learning from underrepresented lesion categories. Each model was trained for a maximum of 100 epochs using a cosine annealing learning rate schedule with a five epoch warm up period. Early stopping was activated when the validation macro F1 score failed to improve for 15 consecutive epochs, and the checkpoint with the highest validation macro F1 score was retained for final testing. The complete training procedure was repeated using random seeds 0, 1, 2, 3, and 4, with identical dataset partitions, preprocessing operations, augmentation policies, and evaluation criteria applied to every model.

4. RESULTS

4.1 Comparison with Established Skin Lesion Classifiers

The comparative evaluation demonstrated a consistent performance advantage for the proposed fusion model across all reported measures, as presented in Table 6. ResNet50 produced the lowest overall performance, with an accuracy of 91.82%, balanced accuracy of 83.45%, macro precision of 85.21%, macro recall of 83.45%, macro F1 score of 84.18%, and macro AUC of 96.13%. DenseNet121 improved these values to 92.48% accuracy, 85.17% balanced accuracy, 86.33% macro precision, 85.17% macro recall, 85.61% macro F1 score, and 96.84% macro AUC, indicating that dense feature reuse provided a moderate improvement over conventional residual learning. EfficientNet B0 achieved a further increase, reaching 93.21% accuracy, 86.14% balanced accuracy, 87.62% macro precision, 86.14% macro recall, 86.73% macro F1 score, and 97.24% macro AUC. Among the single backbone models, Swin Tiny and ConvNeXt Tiny obtained the strongest results. Swin Tiny achieved 94.08% accuracy and 88.69% macro F1 score, whereas ConvNeXt Tiny achieved 94.41% accuracy and 89.37% macro F1 score. The slightly stronger performance of ConvNeXt Tiny suggests that convolutional spatial inductive bias remained beneficial for representing local dermatoscopic texture, pigmentation, and boundary information. Adding pyramid pooling to ConvNeXt increased accuracy to 95.01%, balanced accuracy to 90.47%, macro F1 score to 90.79%, and macro AUC to 98.52%, demonstrating the value of multiscale contextual aggregation. The proposed fusion model achieved the best overall results, with 95.81% accuracy, 92.51% balanced accuracy, 92.64% macro precision, 92.51% macro recall, 92.58% macro F1 score, and 99.01% macro AUC. Compared with ConvNeXt Tiny, the proposed model improved accuracy by 1.40 percentage points, balanced accuracy by 3.67 percentage points, macro F1 score by 3.21 percentage points, and macro AUC by 0.93 percentage points. Relative to ConvNeXt with pyramid pooling, the proposed model produced gains of 0.80 percentage points in accuracy, 2.04 percentage points in balanced accuracy, 1.79 percentage points in macro F1 score, and 0.49 percentage points in macro AUC. The larger improvements in balanced accuracy and macro F1 than in overall accuracy indicate that the proposed fusion strategy improved recognition across the minority categories rather than only increasing predictions for the dominant NV class. Overall, the results in Table 6 show that integrating local convolutional representations, multiscale contextual features, hierarchical transformer features, and gated fusion produced a more balanced classification outcome than any individual baseline architecture.

Table 6. Comparison with established skin lesion classifiers

Model	Accuracy	Balanced accuracy	Macro precision	Macro recall	Macro F1	Macro AUC
ResNet50	91.82%	83.45%	85.21%	83.45%	84.18%	96.13%
DenseNet121	92.48%	85.17%	86.33%	85.17%	85.61%	96.84%
EfficientNet B0	93.21%	86.14%	87.62%	86.14%	86.73%	97.24%
Swin Tiny	94.08%	88.12%	89.43%	88.12%	88.69%	97.91%
ConvNeXt Tiny	94.41%	88.84%	90.11%	88.84%	89.37%	98.08%
ConvNeXt with pyramid pooling	95.01%	90.47%	91.22%	90.47%	90.79%	98.52%
Proposed fusion model	95.81%	92.51%	92.64%	92.51%	92.58%	99.01%

4.2 *Classwise Performance of the Proposed Fusion Model*

The classwise evaluation provides a detailed view of the proposed model performance across the seven diagnostic categories, as shown in Table 7. For AKIEC, the model achieved 89.80% precision, 89.80% sensitivity, 99.66% specificity, an F1 score of 89.80%, and an AUC of 98.71%. These results indicate that the model identified most AKIEC cases while maintaining a low false positive rate. For BCC, precision and sensitivity both reached 93.51%, with a specificity of 99.65%, an F1 score of 93.51%, and an AUC of 99.18%. The strong performance for BCC suggests that the fused representation effectively captured lesion characteristics associated with basal cell carcinoma. The BKL category achieved 90.96% precision, 91.52% sensitivity, 98.88% specificity, a 91.24% F1 score, and a 98.64% AUC. Although BKL contains considerable visual variability and may resemble both melanocytic and keratinocytic lesions, the model maintained balanced precision and sensitivity. For DF, which was represented by only 18 test images, the model achieved 88.89% precision, 88.89% sensitivity, 99.87% specificity, an 88.89% F1 score, and a 98.43% AUC. The comparatively lower F1 score for DF may reflect the limited number of available training and testing examples. For MEL, the proposed model obtained 92.07% precision, 90.42% sensitivity, 99.03% specificity, a 91.24% F1 score, and a 99.12% AUC. These results are important because melanoma shares several visual characteristics with benign melanocytic lesions and represents one of the more difficult classes in the dataset. The dominant NV class achieved the highest number of correct predictions, with 97.82% precision, 98.01% sensitivity, 95.58% specificity, a 97.91% F1 score, and a 99.47% AUC. Its lower specificity relative to the minority classes reflects the larger number of non NV predictions involved in the one versus rest calculation. VASC achieved 95.45% precision, 95.45% sensitivity, 99.93% specificity, a 95.45% F1 score, and the highest classwise AUC of 99.54%. Across all seven categories, the macro precision, macro sensitivity, macro specificity, macro F1 score, and macro AUC were 92.64%, 92.51%, 98.94%, 92.58%, and 99.01%, respectively. The weighted precision, sensitivity, specificity, F1 score, and AUC were 95.80%, 95.81%, 96.78%, 95.80%, and 99.29%, respectively. The difference between the macro and weighted measures reflects the substantial class imbalance in HAM10000, particularly the large contribution of NV to the weighted averages. Nevertheless, macro precision and macro sensitivity above 92% demonstrate that the proposed model maintained relatively stable performance across both majority and minority categories. The results in Table 7 therefore indicate that the model did not achieve its overall accuracy solely through correct classification of NV images, but also provided competitive recognition of less represented classes such as AKIEC, DF, and VASC.

Table 7. Classwise performance of the proposed fusion model

Class	Test images	Precision	Sensitivity	Specificity	F1 score	AUC
AKIEC	49	89.80%	89.80%	99.66%	89.80%	98.71%
BCC	77	93.51%	93.51%	99.65%	93.51%	99.18%
BKL	165	90.96%	91.52%	98.88%	91.24%	98.64%
DF	18	88.89%	88.89%	99.87%	88.89%	98.43%
MEL	167	92.07%	90.42%	99.03%	91.24%	99.12%
NV	1,005	97.82%	98.01%	95.78%	97.91%	99.47%
VASC	22	95.45%	95.45%	99.93%	95.45%	99.54%
Macro average	1,503	92.64%	92.51%	98.97%	92.58%	99.01%

Class	Test images	Precision	Sensitivity	Specificity	F1 score	AUC
Weighted average	1,503	95.80%	95.81%	96.83%	95.80%	99.24%

4.3 Ablation Analysis

The ablation experiments examined the individual and combined contributions of pyramid pooling, Swin Transformer features, direct concatenation, channel attention, and gated fusion, as reported in Table 8. The ConvNeXt Tiny backbone served as the reference configuration and achieved 94.41% accuracy, 88.84% balanced accuracy, 89.37% macro F1 score, and 98.08% macro AUC with 27.8 million parameters. Introducing the pyramid pooling context module increased the parameter count to 29.2 million while improving accuracy to 95.01%, balanced accuracy to 90.47%, macro F1 score to 90.79%, and macro AUC to 98.52%. This corresponds to gains of 0.60 percentage points in accuracy, 1.63 percentage points in balanced accuracy, 1.42 percentage points in macro F1 score, and 0.44 percentage points in macro AUC, indicating that multiscale contextual information contributed particularly to balanced recognition across the diagnostic categories. Combining ConvNeXt with Swin Transformer features produced 95.28% accuracy, 91.16% balanced accuracy, 91.42% macro F1 score, and 98.67% macro AUC with 55.3 million parameters. Relative to the ConvNeXt Tiny backbone, this configuration improved balanced accuracy by 2.32 percentage points and macro F1 score by 2.05 percentage points, supporting the complementary role of hierarchical transformer features. When ConvNeXt, pyramid pooling, and Swin representations were combined using direct concatenation, performance increased to 95.48% accuracy, 91.74% balanced accuracy, 91.91% macro F1 score, and 98.81% macro AUC. This result shows that combining all three feature sources was beneficial even without adaptive weighting. Replacing direct concatenation with channel attention further improved accuracy to 95.61%, balanced accuracy to 92.04%, macro F1 score to 92.17%, and macro AUC to 98.90%, suggesting that feature recalibration helped emphasize informative channels and suppress redundant responses. The complete model with gated fusion achieved the highest values of 95.81% accuracy, 92.51% balanced accuracy, 92.58% macro F1 score, and 99.01% macro AUC with 57.1 million parameters. Compared with the channel attention configuration, gated fusion contributed an additional 0.20 percentage points in accuracy, 0.47 percentage points in balanced accuracy, 0.41 percentage points in macro F1 score, and 0.11 percentage points in macro AUC while adding only 0.3 million parameters. Relative to the original ConvNeXt Tiny backbone, the complete model improved accuracy by 1.40 percentage points, balanced accuracy by 3.67 percentage points, macro F1 score by 3.21 percentage points, and macro AUC by 0.93 percentage points. The progressive improvements observed across the configurations indicate that the final performance was not attributable to a single module. Instead, pyramid pooling improved multiscale context representation, Swin features strengthened long range spatial modeling, channel attention refined feature selection, and gated fusion provided adaptive integration of the three complementary representations. The ablation findings in Table 8 therefore support the contribution of each major architectural component to the final classification performance.

Table 8. Ablation analysis of the proposed architecture

Configuration	Accuracy	Balanced accuracy	Macro F1	Macro AUC	Parameters
ConvNeXt Tiny backbone	94.41%	88.84%	89.37%	98.08%	27.8 M
ConvNeXt with pyramid pooling	95.01%	90.47%	90.79%	98.52%	29.2 M
ConvNeXt with Swin features	95.28%	91.16%	91.42%	98.67%	55.3 M
ConvNeXt, pyramid pooling, Swin, and direct concatenation	95.48%	91.74%	91.91%	98.81%	56.5 M
ConvNeXt, pyramid pooling, Swin, and channel attention	95.61%	92.04%	92.17%	98.90%	56.8 M
Complete model with gated fusion	95.81%	92.51%	92.58%	99.01%	57.1 M

5. DISCUSSION

The experimental findings indicate that combining local convolutional features, multiscale contextual information, and hierarchical transformer representations can improve multiclass classification of dermatoscopic images. The proposed fusion model achieved an accuracy of 95.81%, balanced accuracy of 92.51%, macro F1 score of 92.58%, and macro AUC of 99.01%. Although overall accuracy was high, the balanced accuracy and macro F1 score provide more informative evidence because the HAM10000 dataset is strongly dominated by the NV category. The relatively close values of macro precision, macro recall, and macro F1 suggest that the model maintained a reasonable balance between false positive and false negative predictions across the seven diagnostic categories. The comparison with established classifiers showed progressive improvements from conventional convolutional networks to more advanced contextual and transformer based architectures. ResNet50, DenseNet121, and EfficientNet B0 achieved lower balanced accuracy and macro F1 scores than Swin Tiny and ConvNeXt Tiny. This pattern suggests that modern hierarchical representation learning was more effective for capturing the complex visual characteristics of skin lesions. ConvNeXt Tiny slightly outperformed Swin Tiny, which may indicate that convolutional spatial inductive bias remains particularly useful for dermatoscopic images. Fine pigmentation structures, lesion borders, vascular patterns, and local texture variations are often spatially concentrated, and these characteristics can be represented effectively through hierarchical convolutional operations. The addition of pyramid pooling improved the ConvNeXt Tiny macro F1 score from 89.37% to 90.79%. This improvement supports the importance of examining lesion features at multiple spatial scales. Dermatoscopic abnormalities may appear as small localized structures, broad pigmentation regions, irregular borders, or patterns distributed across the complete lesion. A representation derived from only the final local feature map may therefore overlook relationships between regional and global structures. Pyramid pooling addressed this limitation by combining information obtained from pooling bins of 1, 2, 3, and 6. The resulting context representation complemented the local ConvNeXt features and produced stronger balanced performance without requiring several independently resized input images.

The Swin Transformer branch provided a further improvement by modeling hierarchical relationships between spatially separated lesion regions. The ConvNeXt and Swin configuration achieved a macro F1 score of 91.42%, which was 2.05 percentage points higher than the ConvNeXt Tiny backbone. This finding suggests that attention based global modeling contributed information that was not fully represented by convolutional processing alone. In dermatoscopic images, diagnostic interpretation can depend on relationships among lesion centers, borders, peripheral structures, pigmentation distributions, and surrounding skin. Shifted

window attention enables the gradual exchange of information between nearby and distant regions while preserving hierarchical spatial organization.

The complete three branch configuration achieved stronger performance than either the pyramid pooling or Swin enhanced configurations. Direct concatenation of ConvNeXt, pyramid pooling, and Swin features produced a macro F1 score of 91.91%, confirming that the three feature sources contained complementary information. Channel attention increased this value to 92.17%, while gated fusion further improved it to 92.58%. These results indicate that simply combining heterogeneous representations was beneficial, but adaptively controlling their relative contributions was more effective. The importance of local texture, multiscale context, and global spatial organization may vary among lesion images. Gated fusion allows the model to adjust the influence of each representation rather than assigning equal importance to every branch.The class specific results provide additional evidence regarding the behavior of the proposed model. The NV category achieved the highest F1 score of 97.91%, which was expected because it contained most of the available training images. However, strong results were also observed for several minority categories. VASC achieved an F1 score of 95.45%, while BCC achieved 93.51%. AKIEC and DF produced lower F1 scores of 89.80% and 88.89%, respectively. The performance for DF should be interpreted cautiously because only 18 DF images were included in the test set. With such limited support, a small number of prediction errors can substantially affect sensitivity, precision, and F1 score. The small number of VASC samples introduces a similar concern, even though the reported performance was high.

Melanoma classification remains particularly important because melanoma may share visual characteristics with benign melanocytic lesions. The proposed model obtained 92.07% precision, 90.42% sensitivity, a 91.24% F1 score, and a 99.12% AUC for MEL. The difference between AUC and sensitivity indicates that the model generally ranked melanoma cases effectively, although the final decision threshold still resulted in some false negative predictions. This distinction is important because AUC evaluates discrimination across a range of thresholds, whereas sensitivity reflects performance at the selected operating point. Future evaluation should therefore investigate threshold selection, calibration, and clinically motivated sensitivity requirements rather than relying exclusively on the default maximum probability decision rule.

The proposed model required approximately 57.1 million parameters, compared with 27.8 million parameters for the ConvNeXt Tiny baseline. The additional computational cost resulted from the Swin Transformer branch, pyramid pooling module, projection layers, and fusion mechanism. The gain of 3.21 percentage points in macro F1 suggests that the additional components improved balanced recognition, but the increased complexity may limit deployment on devices with restricted memory or processing capacity. Future work should evaluate parameter reduction, knowledge distillation, pruning, quantization, and lightweight transformer designs. Inference time, computational operations, peak memory use, and energy requirements should also be reported before making claims regarding practical deployment. Several limitations should be acknowledged. First, the experiments were conducted on a single public dataset, and the resulting performance may not represent generalization to other populations, imaging devices, or clinical settings. Second, the rare diagnostic categories contained relatively few samples, which limits the statistical certainty of their class specific results. Third, the current framework performs image level classification and does not incorporate patient history, anatomical location, age, sex, or other clinical variables that may influence diagnosis. Fourth, the model does not provide a direct assessment of predictive uncertainty or probability calibration. Finally, although

the fusion architecture improved classification metrics, further interpretability analysis is required to determine whether the model focuses on clinically meaningful lesion regions.

6. LIMITATIONS AND FUTURE WORK

6.1 Limitations

Several limitations should be considered when interpreting the findings of this study. First, the proposed framework was evaluated only on the HAM10000 dataset. Although this dataset contains images collected from multiple populations and acquisition systems, performance on a single public dataset cannot confirm generalization across different clinical institutions, imaging devices, skin tones, geographic populations, and acquisition protocols. External validation is therefore required before the model can be considered reliable for broader clinical environments. Second, the dataset presents severe class imbalance, with melanocytic nevi representing most available images, while dermatofibroma and vascular lesions contain relatively few samples. Consequently, the class specific results for rare categories are based on small test groups and may be sensitive to only a few incorrect predictions.

The current framework also performs image level classification without incorporating clinical information such as patient age, sex, anatomical location, lesion history, symptoms, or previous diagnostic findings. These variables may provide important evidence during clinical decision making. In addition, the study did not evaluate probability calibration, predictive uncertainty, or selective prediction, which limits the ability to identify cases in which the model is uncertain. Although the proposed architecture combines complementary feature representations, its approximately 57.1 million parameters increase computational requirements compared with the ConvNeXt Tiny baseline. Inference latency, memory consumption, energy requirements, and performance on resource constrained devices were not examined. The present study also did not include detailed interpretability analysis to verify whether model predictions were based on clinically meaningful lesion structures rather than surrounding artifacts, acquisition markers, or background patterns. The reported results should additionally be interpreted within the experimental design. While lesion level partitioning reduced information leakage between related images, it did not explicitly control for possible patient, clinical center, acquisition device, or source based overlap. Statistical comparisons between the proposed model and the baseline classifiers were not reported, and the practical importance of relatively small performance differences should therefore be interpreted cautiously. The model was trained using a fixed input resolution of 224×224 pixels, which may reduce the visibility of very small structures in high resolution dermatoscopic images. Finally, the framework predicts one of seven predefined diagnostic categories and does not address lesion segmentation, unknown lesion types, poor quality images, or cases containing multiple clinically relevant abnormalities.

6.2 Future Work

Future research should evaluate the proposed framework on independent dermatoscopic datasets collected from different institutions, populations, imaging devices, and skin tone groups. Cross dataset testing and external validation would provide stronger evidence regarding model robustness and generalization. Prospective clinical evaluation should also be considered to examine performance under routine acquisition conditions. Particular attention should be given to rare lesion categories through larger multicenter

datasets, balanced sampling, class aware augmentation, generative data augmentation, and few sample learning strategies. Confidence intervals and statistical comparison methods should be included to determine whether observed improvements remain stable across experimental repetitions. The framework can also be extended by integrating patient metadata and relevant clinical variables with the learned image representation. Multimodal fusion may improve discrimination between visually similar lesion categories and provide predictions that better reflect clinical reasoning. Future versions should incorporate probability calibration and uncertainty estimation to identify ambiguous cases and support referral to human specialists. Selective prediction mechanisms may allow the model to provide predictions only when confidence is sufficiently reliable while abstaining from uncertain cases.

Further investigation should focus on interpretability and computational efficiency. Attention visualization, gradient based localization, concept based explanations, and expert evaluation could be used to determine whether the model focuses on diagnostically meaningful lesion regions. Model compression through knowledge distillation, pruning, quantization, parameter sharing, and lightweight transformer components may reduce computational cost while preserving balanced classification performance. Future studies may also explore multiresolution input processing, lesion guided attention, automatic segmentation, artifact removal, and open set recognition. These extensions could improve robustness to image quality variation, support recognition of previously unseen lesion categories, and strengthen the potential value of the framework as an image analysis support system.

7. CONCLUSION

This study presented a hybrid deep learning framework for seven class dermatoscopic image classification by integrating ConvNeXt Tiny, pyramid pooling, Swin Transformer, and gated feature fusion. The ConvNeXt branch captured local texture, color, border, and morphological information, while the pyramid pooling module aggregated lesion context at multiple spatial scales. The Swin Transformer branch modeled hierarchical spatial relationships and broader lesion dependencies. These complementary representations were projected into a shared feature space and adaptively combined before final classification.On the lesion level HAM10000 partition, the proposed model achieved 95.81% accuracy, 92.51% balanced accuracy, 92.64% macro precision, 92.51% macro recall, 92.58% macro F1 score, and 99.01% macro AUC. Comparative and ablation results indicated that pyramid pooling, transformer based global modeling, channel attention, and gated fusion each contributed to the final performance. The larger improvement in balanced accuracy and macro F1 compared with overall accuracy suggests that the framework improved recognition across both common and underrepresented lesion categories. The findings demonstrate the potential value of combining local, contextual, and global visual representations for imbalanced skin lesion classification. However, the proposed framework should be considered an image analysis support method rather than an autonomous diagnostic system. External validation, statistical evaluation, calibration analysis, uncertainty estimation, interpretability assessment, and computational optimization remain necessary before broader clinical applicability can be established.

Author Contributions: Sumayya Jahan conceived the study, developed the research methodology, and coordinated the overall investigation. Kallol Chakraborty Shekhor contributed to the architectural design, literature review, and preparation of

the manuscript. Md Sahid Hossain supported model implementation, software development, and experimental analysis. Md Fakrul Alam contributed to data preprocessing, technical validation, and interpretation of the findings. Abdur Rahim assisted with dataset preparation, lesion level partitioning, and performance evaluation. Md Sazzad Hossain Shehan contributed to statistical analysis, result visualization, and manuscript drafting. Md Anwar Hossain and Mohammed Adnan provided technical supervision, reviewed the experimental design, and critically revised the manuscript. Md Abedur Rahman contributed to result verification, manuscript editing, and final quality assessment. All authors reviewed and approved the final version of the manuscript.

Data Availability: The datasets used in this study are publicly available through Kaggle.

Declarations

Ethical approval: Not applicable.

Consent to participate: Not applicable. The study used only secondary, anonymized image data and involved no direct contact or intervention with individual participants.

Consent to publish: Not applicable. This manuscript does not contain any individual person's identifiable data (including names, facial photographs, or other personal details).

Clinical Trial: Not applicable.

Funding: Not applicable.

REFERENCES

[1] Das, K., Cockerell, C. J., Patil, A., Pietkiewicz, P., Giulini, M., Grabbe, S., and Goldust, M. (2021). Machine learning and its application in skin cancer. International Journal of Environmental Research and Public Health, 18(24), 13409.

[2] Manik, M. M. T. G. (2022). An analysis of cervical cancer using the application of AI and machine learning. Journal of Medical and Health Studies, 3(2), 67–76. https://doi.org/10.32996/jmhs.2022.3.2.11

[3] Dildar, M., Akram, S., Irfan, M., Khan, H. U., Ramzan, M., Mahmood, A. R., Alsaiari, S. A., Saeed, A. H. M., Alraddadi, M. O., and Mahnashi, M. H. (2021). Skin cancer detection: A review using deep learning techniques. International Journal of Environmental Research and Public Health, 18(10), 5479.

[4] Manik, M. M. T. G., Saimon, A. S. M., Miah, M. A., Ahmed, M. K., Khair, F. B., Moniruzzaman, M., and Bhuiyan, M. M. R. (2021). Leveraging AI powered predictive analytics for early detection of chronic diseases: A data driven approach to personalized medicine. Nanotechnology Perceptions, 17(3), 269–288. https://doi.org/10.62441/nano-ntp.v17i3.5444

[5] Goceri, E. (2021). Automated skin cancer detection: Where we are and the way to the future. In Proceedings of the 44th International Conference on Telecommunications and Signal Processing, 48–51. IEEE.

[6] Nahata, H., and Singh, S. P. (2020). Deep learning solutions for skin cancer detection and diagnosis. In Machine Learning with Health Care Perspective: Machine Learning and Healthcare, 159–182. Springer International Publishing.

[7] Shekhor, K. C., Hossain, M. S., Rahman, M. A., and Hossain, M. A. Multi Scale Attention Fusion Classification with an LLM Driven Clinical Recommendation Assistant for Brain Tumor MRI. Journal of Medical and Health Studies, 4, 121–139.

[8] Manik, M. M. T. G. (2021). Multi omics system based on predictive analysis with AI driven models for Parkinson's disease neurosurgery. Journal of Medical and Health Studies, 2(1), 42–52. https://doi.org/10.32996/jmhs.2021.2.1.5

[9] Hossain, M. S., Shekhor, K. C., Hossain, M. A., and Rahman, M. A. Diagnosis Everywhere: Lightweight AI That Detects Disease from Scans and Records on Limited Hospital Hardware. Frontiers in Computer Science and Artificial Intelligence, 2, 82–94.

[10] Manik, M. M. T. G. (2023). Multi omics integration with machine learning for early detection of ischemic stroke through biomarkers discovery. Journal of Ecohumanism, 2(2), 175–187. https://doi.org/10.62754/joe.v2i2.6800

[11] Monika, M., Vignesh, N. A., Kumari, C. U., Kumar, M. N. V. S. S., and Lydia, E. L. (2020). Skin cancer detection and classification using machine learning. Materials Today: Proceedings, 33, 4266–4270.

[12] M. M. T. G., Hossain, S., Ahmed, M. K., Rozario, E., Miah, M. A., Moniruzzaman, M., and Muhammad, A. S. (2022). Integrating genomic data and machine learning to advance precision oncology and targeted cancer therapies. Nanotechnology Perceptions, 18(2), 219–243. https://doi.org/10.62441/nano-ntp.v18i2.5443

[13] Vidya, M., and Karki, M. V. (2020). Skin cancer detection using machine learning techniques. In Proceedings of the IEEE International Conference on Electronics, Computing and Communication Technologies, 1–5. IEEE.

[14] Zghal, N. S., and Derbel, N. (2020). Melanoma skin cancer detection based on image processing. Current Medical Imaging Reviews, 16(1), 50–58.

[15] Kumar, M., Alshehri, M., AlGhamdi, R., Sharma, P., and Deep, V. (2020). A DE ANN inspired skin cancer detection approach using fuzzy c means clustering. Mobile Networks and Applications, 25(4), 1319–1329.

[16] Gouda, W., Sama, N. U., Al Waakid, G., Humayun, M., and Jhanjhi, N. Z. (2022). Detection of skin cancer based on skin lesion images using deep learning. Healthcare, 10(7), 1183.

[17] Miah, M. A., Rozario, E., Khair, F. B., Ahmed, M. K., Bhuiyan, M. M. R., and Manik, M. M. T. G. (2019). Harnessing wearable health data and deep learning algorithms for real time cardiovascular disease monitoring and prevention. Nanotechnology Perceptions, 15(3), 326–349. https://doi.org/10.62441/nano-ntp.v15i3.5278

[18] Ameri, A. (2020). A deep learning approach to skin cancer detection in dermoscopy images. Journal of Biomedical Physics and Engineering, 10(6), 801.

[19] Thurnhofer Hemsi, K., and Domínguez, E. (2021). A convolutional neural network framework for accurate skin cancer detection. Neural Processing Letters, 53(5), 3073–3093.

[20] M. M. T. G., Moniruzzaman, M., Islam, M. S., Bhuiyan, M. M. R., Rozario, E., Hossain, S., and Saimon, A. S. M. (2020). The role of big data in combatting antibiotic resistance: Predictive models for global surveillance. Nanotechnology Perceptions, 16(3), 361–378. https://doi.org/10.62441/nano-ntp.v16i3.5445

[21] Rajarajeswari, S., Prassanna, J., Quadir, M. A., Jackson, J. C., Sharma, S., and Rajesh, B. (2022). Skin cancer detection using deep learning. Research Journal of Pharmacy and Technology, 15(10), 4519–4525.

[22] Reis, H. C., Turk, V., Khoshelham, K., and Kaya, S. (2022). InSiNet: A deep convolutional approach to skin cancer detection and segmentation. Medical and Biological Engineering and Computing, 60(3), 643–662.

[23] T. G., Bhuiyan, M. M. R., Moniruzzaman, M., Islam, M. S., Hossain, S., and Hossain, S. (2018). The future of drug discovery utilizing generative AI and big data analytics for accelerating pharmaceutical innovations. Nanotechnology Perceptions, 14(3), 120–135. https://doi.org/10.62441/nano-ntp.v14i3.4766

[24] Imran, A., Nasir, A., Bilal, M., Sun, G., Alzahrani, A., and Almuhaimeed, A. (2022). Skin cancer detection using combined decision of deep learners. IEEE Access, 10, 118198–118212.

[25] Manik. (2020). Biotech driven innovation in drug discovery: Strategic models for competitive advantage in the global pharmaceutical market. Journal of Computational Analysis and Applications, 28(6), 41–47. https://eudoxuspress.com/index.php/pub/article/view/2874

[26] Pacheco, A. G., and Krohling, R. A. (2020). The impact of patient clinical information on automated skin cancer detection. Computers in Biology and Medicine, 116, 103545.

[27] Chowdhury, M. S. I., Chy, K. S., Islam, M. A., & Chowdhury, M. N. I. (2023). Integrating Blockchain and Data Analytics to Strengthen Financial Traceability and Anti-Fraud Controls. *Journal of Computer Science and Technology Studies*, *5*(3), 189-202. https://doi.org/10.32996/jcsts.2023.5.3.14

[28] Chowdhury, M. N. I., Chy, K. S., Chowdhury, M. S. I., & Islam, M. A. (2021). Identifying Hidden Fraud Indicators through Longitudinal Analysis of U.S. Financial Misconduct. Journal of Economics, Finance and Accounting Studies, 3(1), 89-99. https://doi.org/10.32996/jefas.2021.3.1.9

[29] Chy, K. S., Islam, M. A., Chowdhury, M. S. I., Chowdhury, M. N. I., & Islam, M. A. (2022). Enhancing U.S. Financial Compliance and Risk Management through Data-Driven Automation and Anomaly Detection. Journal of Business and Management Studies, 4(3), 231-244. https://doi.org/10.32996/jbms.2022.4.3.22