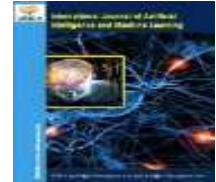




International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

A Hybrid Distortion-Robust Skin Disease Classification Using Generative Preprocessing and Cross-Level CNN-Transformer Fusion

Bipin P. R¹, Sajin C S², Anoop V³, Prasanth P Menon⁴, Upendra Kumar⁵, Sai Kiran Oruganti⁶

¹Post Doctoral Research Fellow, Lincoln University College, Malaysia. Email: pdf.bipin@lincoln.edu.my

²Department of Electronics and Communication Engineering, Adi Shankara Institute of Engineering and Technology, Kalady, Kerala, India, Email: sujo.er@gmail.com

³Department of Artificial Intelligence and Data Science, Jyothi Engineering College, Thrissur, Kerala 679531, India. E-mail: vanoop@gmail.com

⁴Department of Electronics and Communication Engineering, Adi Shankara Institute of Engineering and Technology, Kalady, Kerala, India. E-mail: prasanthpraveenmenon@gmail.com

⁵Institute of Engineering and Technology, UP, India. E-mail: upendra.kumer.lincoln@gmail.com

⁶Professor, Lincoln University College, Malaysia. E-mail: saikiran.lincoln@gmail.com

*Corresponding authors: Bipin P R, Sai Kiran Oruganti

Abstract:

Skin cancer is one of the most serious forms of cancer and melanoma is a very serious form if diagnosed too late. Although dermatoscopy is a non-invasive diagnostic tool for visualizing skin lesions, its diagnostic reliability is influenced by variability in image quality, overlapping visual features of the classes of disease diagnosis, and clinical interpretation subjectivity. In recent years, deep neural networks, particularly Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), have shown promising results in assisting with the automatic detection and categorization of dermatological abnormalities. However, CNNs are limited in their ability to model global dependencies and ViTs' ability to generalize is limited by the amount of data. Furthermore, many existing methods are mainly evaluated against clean dermoscopic images, and do not explicitly consider robustness with respect to different image distortions.

To address these deficiencies, the current study introduces a hybrid framework combining generative preprocessing with CNN-Transformer cross-level fusion, which is robust to distortion. A generative adversarial network (GAN) is used to denoise, reduce blur and regularize illumination inconsistencies, while boosting the saliency of lesions. The CNN branch is used to extract local features and the ViT branch is employed to capture global relationship, then the two outputs are fused through cross-level fusion. The proposed approach outperforms the CNN-only, ViT-only and conventional hybrid approaches as shown by the evaluations performed on the HAM10000 dataset under clean and distorted conditions, and external cross-dataset validation on ISIC dataset. This work gives a major boost to the robustness of skin disease classification and provides a potential avenue for AI in dermatology.

1. Introduction

Skin cancer is becoming a major health concern and an increasing problem in the world, and has important implications for public health. Epidemiological surveys show that the incidence of melanoma is on the rise in recent years, especially among fair-skinned people. By far, the most common types of skin cancer are non-melanoma, but the deadliest type is melanoma because it has a strong likelihood of spreading. Early detection is therefore paramount as survival is significantly increased when the disease is detected at its early stages.

Dermatoscopy, which helps to capture the objects beneath the skin, is becoming an everyday tool for dermatologists. It enhances the visualization of hidden lesion characteristics not normally discernible by direct observation. Though dermatoscopic diagnosis is a valuable tool, there is some variation in its accuracy and degree of accuracy depends on the experience of the clinician, their expertise and their consistency in interpreting the visual clues. In order to reduce this subjectivity, research has been focusing on computer-aided diagnostic systems, specifically deep learning-based ones [1,2,3]. These approaches can train on a large number of samples to learn discriminative information and make objective predictions to assist in clinical decisions.

1.1 Applications of CNN for Skin Disease Localization

CNNs have been the highlights of computer vision for quite a while now. They are designed to take advantage of the spatial locality of images. In dermatology, the CNN-based lesion detection, segmentation and classification have been applied extensively [1,2,4]. Networks like ResNet, DenseNet, and Inception have achieved results close to and even superior to dermatologists' accuracy on benchmark datasets like ISIC and HAM10000.

CNNs have several benefits: They are efficient, well-studied, and they can learn hierarchies of features at multiple levels. Early layers tend to represent low level features such as pigmentation or edge sharpness,

and deeper layers tend to represent high level structures. However, CNNs have a natural roof on their receptive field size, which makes it challenging to capture the global context relations without switching to deeper and more complex models. This limitation makes classification less reliable in dermatology, where the overall distribution of the lesion pattern and the structure of the image in the global image plane can be as critical as the local edges.

1.2 Emergence of Vision Transformers

Vision Transformers (ViTs) are different than conventional convolution models as they propose a new approach to image understanding [5]. They segment an image applied to it into predetermined patches, converting them into an embedding, and use self-attention to model interactions between far-apart regions. This means that when used to model global dependencies, ViTs are naturally good at this, whereas CNNs are less capable of such tasks where the entire structure of the image is important.

The viability of ViTs in medical imaging, such as dermatology, has started to become demonstrated. They can focus on specific regions of the image which are located in different areas, and detect small variations in color and texture, that CNNs may not detect. There are however challenges: ViTs need large training sets to generalize, and many medical imaging datasets are quite small and imbalanced. In addition, the ViTs are sensitive to noise and artefacts that can lead to mis attention to irrelevant areas [6,7].

1.3 Hybrid CNN-Transformer Architectures

Given the complementary strengths of CNNs and ViTs, there has been a start of designing hybrid models. CNNs are good at learning strong local features, while ViTs can capture long-range dependencies. Now when they are linked together, one can get the best of both worlds.

Early CNN-ViT hybrids usually involved late fusion approaches, where the features learned by the CNN and ViT were simply concatenated and passed through a classification head. These methods achieved better results than single models, but typically suboptimal leveraging of the hierarchical modeling capabilities of CNNs and multi-level modeling capabilities of ViTs. In addition, the majority of the hybrids examined so far have been developed and tested on dermoscopic images in which the presence of distortions has not been accounted for.

1.4 Distortion as a Practical Challenge

In real-world practice, dermatoscopic images are seldom ideal. There are many possible types of artifacts and distortions, including gaussian noise, motion blur due to camera shakes or subject motion, shadows and intensity variations due to uneven lighting.

To avoid these challenges pre-cleaned datasets are used in the literature, but impractical in real-world clinical workflows. A powerful computer aided diagnostic system must be able to perform properly even in the presence of distortion. This need highlights the need for strong preprocessing techniques.

1.5 Generative Preprocessing for Robustness

Standard techniques like blur reduction or histogram equalization are not flexible and are often insufficient to process the wide range of dermatoscopy distortions. The idea of generative models, notably GANs, is to bring a data-driven and adaptive approach toward preprocessing. GANs are trained to map the distorted image to its clean version, enabling them to selectively remove noise, artifacts and occlusion while preserving information specific to the lesions. The dynamic enhancement of the image quality makes GANs an appealing option for pre-processing in pipelines designed for skin disease classification.

1.6 Proposed Method

To address these research problems, we propose a CNN-ViT hybrid model with generative pre-processing & cross-level feature fusion, which is robust to distortion. The plan has the following components:

- Generative Preprocessing: A GAN based network is used to synthesize clear and distinct dermatoscopic images from noisy, blurry and illumination distorted images, preserving the structure of the lesion.
- Dual-Path Feature Extraction: The global dependencies are captured by a ViT branch, and the hierarchical local features are extracted by a CNN branch.
- Cross-Level Fusion: Embeddings from several CNN layers are combined with transformer embeddings to provide a holistic representation that uses both local and global information.

1.7 Contributions

The central innovations presented in this research can be highlighted as:

- Introduction of GAN-based generative preprocessing to improve robustness against distortions common in dermatoscopic images.
- Development of a cross-level CNN-ViT integrated framework capable to simultaneously capture intricate local textures and extensive contextual information.
- Comprehensive evaluation under both clean and distorted image conditions, demonstrating robustness not addressed in prior studies.

1.8 Organization of the Paper

In Section 2, literature of CNNs, transformers, hybrid architectures and pre-processing is explored. In Section 3, details of the methodology and preprocessing aspects are discussed. Ablation studies are presented and possible conclusion is drawn in Section 4.

2. Literature Review

In recent years, there has been an exponential increase in the number of research papers related to automated skin disease classification because of rapid progress in deep learning and the creation of dermatoscopic image datasets. A number of methods have been developed such as CNN based models, Vision Transformer based models, and hybrid CNN-ViT models, as well as artifact removal pre-processing methods. This section will detail the review of these works, strengths, weaknesses and gaps that the proposed study will attempt to fill.

2.1 Role of CNNs in Automated Skin Disease Classification

In the field of healthcare, particularly dermatology, CNNs have long been the backbone of computer vision applications. In their groundbreaking study, Esteva et al. [1] showed that a CNN trained using 129,450 skin images, both dermatoscopic and clinical images, can achieve accuracy levels similar to dermatologists and distinguish between malignant and benign skin lesions. This was a ground breaking project that founded the viability of CNNs as diagnostic tools.

Later, deeper architectures were adopted. Shehzad et al. [2] used an ensemble of deep learning models for lesion classification and used the residual connections to demonstrate that convergence and accuracy are improved over shallower CNNs. Likewise, Codella et al. [3] used an ensemble of multiple CNNs to enhance the ability to detect melanoma in ISIC challenge dataset with notable improvements in F1-scores. A few years later, Gessert et al. [4] benchmarked DenseNet, ResNet and EfficientNet variants and found that under certain circumstances, a carefully tuned ensemble of CNNs can outperform more complex ensembles.

Although CNNs have succeeded, they have significant drawbacks. They have wide receptive fields to expand slowly. They need deep networks to model the lesion structures at a global level, which is very difficult. Furthermore, CNNs tend to misclassify both local and global textures are similar, but morphology is different. These drawbacks motivated discussing transformer-based approaches.

2.2 Vision Transformers in Medical Imaging and Dermatology

Transformers were once revolutionary in the field of language modelling and have similarly demonstrated their impact in computer vision tasks. ViT (Dosovitskiy et al. [5]) was proposed as an extension of the work that replaces CNN with self-attention layers using an image as an ordered sequence of image segments (patches) and thus captures global relationships better than CNNs.

In the dermatology field, there is promising research into the use of ViTs. In [6] G M S. Himel et al. used ViT models on the HAM10000 dataset and found that Vision Transformer models are effective for automated skin image classification. They showed that without explicit localization they could focus on clinically relevant regions, such as pigment networks. Similarly, Khan et al. [7] investigated the performance of the transformers on medical imaging applications and concluded that the transformers perform better than CNNs in skin lesion classification with multiple classes.

However, due to their strong data-dependency, ViTs cannot be readily applied to medical imagery because of the scarcity of annotated data. In addition, they are easily affected by noise and distortions, which can make them less robust. The constraints instigated hybrid architectures that fuse CNNs with transformers.

2.3 Hybrid CNN-Transformer Architectures

To take the strengths of both the CNNs and ViTs, hybrid models are developed. To classify skin lesions, Mahbod et al. [8] experimented with fusing features extracted from various CNNs, and found that the combination of features extracted from different algorithms yielded significant improvement in accuracy on the ISIC dataset than using any single algorithm. Likewise, Chen et al. [9] introduced TransUNet for medical segmentation combining CNN encoders with transformer decoders for the localization of organs. Tu et al. [10] proposed MaxViT that uses convolutional blocks along with the local and global attention layers to achieve good ImageNet results. Huang et al. [10] proposed a CNN that removed the necessity of directly modeling spatial relations by introducing a new architecture, ConvNeXt, inspired by the transformer architecture. Liu et al. [11] proposed ConvNeXt, a new CNN inspired by the transformer architecture that narrowed the gap between CNN and transformer performance. To further enhance training stability and efficiency, Woo et al. [12] presented a new network called ConvNeXtV2. Although many hybrid applications are proposed in dermatology, they are mostly based on late feature concatenation, with the exception of high-level feature fusion.

2.4 Preprocessing and Artifact Removal in Dermatoscopic Images

Good preprocessing for reducing irrelevant information is crucial to improving the accuracy of classification models. Traditional methods include hair removal filters [13], illumination correction, and

histogram equalization [14,15]. These are useful but are static and can end up misrepresenting the structures of the lesions.

Learnable preprocessing has been recently investigated. For increase of relevant areas before the classification, Yasir et al. [16] employed attention modules to focus the lesion content. Al-Masni et al. [17] segmented and classified the images using ensemble deep learning techniques, only analyzing the lesion areas. Another promising approach is the GAN-based approach, a preprocessing method. Bipin et al. [18] has suggested an algorithm based on Hybrid feature refinement using CNN. Halladja et al. [19] used EA-GAN to clean up dermatoscopic images containing hair occlusion, with encouraging results obtained when they were used to classify the images. These results highlight the potential of generative models for adaptive and distortion-tolerant preprocessing.

2.5 Gaps in Existing Literature

The literature reveals several gaps:

- CNNs excel locally but lack global context.
- ViTs provide global features but require large data and are sensitive to distortions.
- Hybrid CNN-ViT models exist but often rely on naive late fusion.
- Preprocessing is mostly traditional; GAN-based adaptive methods are rare.
- Few works evaluate robustness under distortions common in dermatoscopy.

In general, CNNs and ViTs have contributed to the improvement of skin disease classification, and although CNN-ViT based models have been developed, skin disease classification still has a lack of models that can handle both robustness and advanced feature fusion. To remedy these drawbacks, the proposed framework utilizes GAN-based pre-processing and cross-level CNN-ViT fusion, and thereby provides a holistic solution for the domain.

3. Methodology

The deep learning models applied to dermoscopic images have been found to have good accuracy in the ideal settings of the benchmarking data sets, but fail to perform as well when they are used in clinical settings, where the image quality is not always optimal. To overcome this, the proposed system is composed of two parts: (i) the CNN and ViT dual-branch feature extraction backbone and (ii) the GAN-based preprocessing module to improve the quality of dermoscopic images and reduce distortions. These components create an integrated system that enhances the accuracy and reliability of automated dermatological diagnostics.

3.1 Dataset Description

A primary dataset, HAM10000, provides a wide collection of 10,015 dermatoscopic images across seven diagnostic classes (Melanoma, Melanocytic nevus, Basal cell carcinoma, Actinic keratosis, Benign keratosis, Dermatofibroma and Vascular lesion). Each image is annotated by expert comments, which makes the labelling of the ground truth reliable. The HAM10000 dataset was split into three groups: 70% for training, 15% for validation and 15% for testing. The training subset was used to train the GAN, the hybrid classifier, and the validation subset was used to monitor the performance of the model. Further, two types of test conditions were created from the test subset (which was kept separately from the training process), they are:

- Clean Test Set: images without any artificial distortions.
- Distorted Test Set: images with artificial motion blur, Gaussian noise and uneven illumination.

This twofold evaluation approach enables evaluation in clean and in simulated conditions with inevitable imaging defects. For external cross-dataset validation we have employed ISIC dataset, which is to evaluate the generalizability of the proposed framework.

3.2 Preprocessing

Basic issues that can occur with raw dermoscopic images include artifacts and inconsistencies, including occlusion by hair, varying illumination, compression noise, and sensor-related distortion. These artifacts may mask features that are important for diagnosis, such as pigment networks, streaks and lesion asymmetry, and hamper the performance of image-based classifiers. Traditional pre-processing methods include fixed algorithms such as DullRazor for removing hairs and contrast enhancement algorithms such as histogram equalization. These techniques are static and non-adaptive, but they work well in some situations. They may miss out on minor structures of the lesions or add artificial edges to it, which compromises the fidelity.

To overcome these limitations, the present study employs a Generative Adversarial Network (GAN)-based preprocessing module. Unlike fixed rule-based filters, the GAN learns in a data-driven fashion how to enhance image quality by suppressing noise while preserving diagnostically critical structures. This ensures that subsequent feature extraction focuses on clinically meaningful patterns rather than artifacts.

GAN-Based Preprocessing Module:

The GAN-based preprocessing module is central to the robustness of the proposed system. It is designed to transform distorted dermoscopic images into high-quality, enhanced and faithful representations. It is implemented using a conditional GAN with a U-Net generator (Gr) and a PatchGAN discriminator (Dr).

- When distorted image I_{dist} is applied as the input the generator outputs a cleaned version $I_{clean} = Gr(I_{dist})$
- The Discriminator's role is to differentiate genuine clean images I_{real} and generated clean outputs $Gr(I_{dist})$

The GAN is trained on images from HAM10000 training subset only. The generator learns to translate the distorted image back into its clean form. The role of the discriminator is to identify distinctions between actual clean images and the generator's outputs. Training uses a mix of adversarial loss (to ensure realism) and reconstruction loss (L1) (to ensure fidelity to the clean reference). This ensures the GAN does not hallucinate features but focuses on restoring lesion clarity.

Synthetic distortions are created only from the corresponding HAM10000 training images while GAN training. To avoid the data leakage, testing and validation sets are not used for GAN training.

Adversarial Objective:

The adversarial objective ensures that generated outputs look realistic:

$$L_{GAN}(Gr, Dr) = E_{I_{real}} [\log Dr(I_{real})] + E_{I_{dist}} [\log (1 - Dr(Gr(I_{dist}))) \quad (1)$$

The loss described in equation (1) drives the generator Gr to produce outputs that maximize the likelihood of being classified as real by the discriminator Dr.

Reconstruction Loss:

To ensure that enhanced images remain faithful to the true lesion structure, a reconstruction loss is added, shown in equation (2):

$$L_{rec}(Gr) = || I_{real} - Gr(I_{dist}) ||_1 \quad (2)$$

This term penalizes deviations from the ground truth clean image, ensuring that the generator does not hallucinate spurious structures.

Combined Objective:

The total loss for GAN-based preprocessing is a weighted sum:

$$L_{total} = L_{GAN} + \lambda L_{rec}$$

where λ balances realism with lesion fidelity, which is set as 100 here.

The outcome of the GAN preprocessing module is an enhanced image I_{clean} , where gaussian noise and motion blur are suppressed, illumination is normalized, and lesion boundaries are sharpened. These images serve as the input to the CNN and Transformer branches.

3.3 CNN Branch: Local Feature Extraction

After the GAN based pre-processing module, the dermoscopic images are fed to the convolutional branch of the model. CNN models can be effectively applied in the detection of localized structures in images, which is highly relevant in dermatological applications, where the diagnostic information may be found in small details like pigment irregularities, streaks, or sharp changes in texture. We use ConvNeXtV2-Tiny initialized with pretrained ImageNet weights and finetuned on HAM10000 training subset as the CNN backbone, which were found to take advantage of these strengths. This decision is made because we need a modern CNN architecture that can contain low level descriptors as well as provide hierarchical depth and scale similar to the transformer-based approach. The ConvNeXtV2-Tiny model used here uses depth wise convolution, GELU activation and Global Response Normalization.

The CNN branch takes as input the image I_{clean} , enhanced by the GAN, standardized to the size 224x224x3. This image is passed through a series of convolutional layers, each of which produces increasingly abstract features in the image. Formally, the CNN feature map at the l^{th} layer is denoted as, equation (3):

$$F_{CNN}^{(l)} = f_{CNN}^{(l)}(F_{CNN}^{(l-1)}) \quad (3)$$

In this, $F_{CNN}^{(l)}$ represents the output feature map at layer l , $f_{CNN}^{(l)}$ denotes convolution operation at layer l including normalization and activation. $F_{CNN}^{(0)} = I_{clean}$

This recursive notation helps to understand that the output of each layer will depend on the previous layer's output and not on the raw input image.

- Low-level features $F_{CNN}^{(1)}$ contain information about edges, simple colour gradients, and local contrasts, which correspond to the earliest diagnostic cues visible to dermatologists.
- Mid-level features $F_{CNN}^{(2)}$ encode more complex structures, such as pigment networks, dotted patterns, or clustered textures. These are often linked with early pathological changes.
- High-level features $F_{CNN}^{(3)}$ represent lesion-wide morphology, including asymmetry and overall geometric irregularities, which are critical for differentiating malignant melanomas from benign lesions.

3.4 Vision Transformer Branch: Global Context Extraction

While CNNs provide strong local feature extraction, their receptive fields are limited. They often fail to capture long-range dependencies such as lesion symmetry across opposite quadrants or irregular pigment spread across the entire lesion. To complement the CNN branch, the second backbone is a Vision Transformer (ViT), which excels at modeling global dependencies through attention mechanisms.

In this work, we use the ViT-B/16 configuration pretrained with ImageNet weight and finetuned using HAM10000 training subset. The input image of size 224×224 is partitioned into 16×16 non-overlapping patches, giving a total of $N = (224/16)^2 = 196$ patches. Each patch is flattened into a one-dimensional vector x_i (for $i = 1, 2, \dots, N$) and mapped to a 768-dimensional embedding (D). Positional encodings are then incorporated to retain spatial context, forming the input token sequence as shown in (4):

$$z_0 = [x_{cls}; x_1 E; x_2 E; \dots; x_N E] + E_{pos} \quad (4)$$

where, x_i = flattened vector of pixel intensities for the i^{th} patch, E = learnable embedding matrix mapping patch vectors to dimension D , E_{pos} = positional encoding vector that encodes spatial location of each patch and x_{cls} = classification token, a special learnable embedding that aggregates global information.

The embedding sequence z_0 is refined through multiple stages, specifically 12 layers of Transformer encoders. At the k^{th} encoder, the update rule is represented as in equation (5):

$$z_k' = \text{MHSA}(\text{LN}(z_{k-1})) + z_{k-1}, z_k = \text{MLP}(\text{LN}(z_k')) + z_k' \quad (5)$$

where, LN = Layer Normalization, MHSA = Multi-Head Self-Attention, MLP = position-wise feed-forward network.

Multi-Head Self-Attention (MHSA)

Let $Z \in \mathbb{R}^{N \times D}$ represent the matrix of token embeddings at a given layer. Queries (Q), Keys (K), and Values (V) are computed using learnable projection matrices illustrated in (6):

$$Q = Z W_Q, K = Z W_K, V = Z W_V \quad (6)$$

Where, $W_Q, W_K, W_V \in \mathbb{R}^{D \times d_k}$ are learnable projection matrices and d_k = dimensionality of the key vectors. The attention weights are computed based on equation (7):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

In multi-head attention, the operation is repeated across $h=12$ parallel heads, allowing the model to learn diverse relationships such as colour, shape, and texture at the same time. The resulting outputs from all heads are concatenated and linearly transformed back to dimension D .

Thus, MHSA allows each patch embedding to attend to every other patch, enabling the Transformer to model lesion-wide structural dependencies such as asymmetry or irregular pigmentation spread.

The features extracted at the k^{th} encoder are denoted as in equation (8):

$$F_{ViT}^{(k)} = z_k \quad (8)$$

where $F_{ViT}^{(k)}$ represents the global contextual embedding after the k^{th} Transformer layer.

3.5 Cross-Level CNN-ViT Fusion

While CNN and Transformer branches each provide valuable perspectives, relying on one alone is insufficient for robust dermatological classification. CNNs excel in local feature detection, but miss global context. Transformers capture holistic structures, but are prone to overlooking fine detail when data is limited. To overcome these complementary limitations, we employ a cross-level fusion block, shown in Figure 1.

This block allows bidirectional information flow between CNN and ViT features.

CNN → ViT Injection:

CNN mid-level features $F_{CNN}^{(l)}$ (where l denotes the selected mid-level convolutional layer) are projected into the embedding space of the Transformer. This is performed by applying a projection matrix W_p followed by alignment transformation ϕ :

$$\tilde{F}_{ViT}^{(k)} = \phi(W_p F_{CNN}^{(l)}) + F_{ViT}^{(k)} \quad (9)$$

Here, W_p projects CNN feature maps into the same dimensional space as Transformer embeddings, ϕ reshapes the projected CNN features into token form, $\tilde{F}_{ViT}^{(k)}$ denotes enriched Transformer embeddings at encoder depth k .

This injection ensures that the Transformer's global attention is informed by localized CNN cues, enabling fine lesion textures to influence global reasoning.

ViT → CNN Feedback:

Transformer features $F_{ViT}^{(k)}$ are projected back into CNN feature map space using a learnable transformation W_t and reshaping operator ψ , which is illustrated in (10):

$$\tilde{F}_{CNN}^{(l)} = \psi(W_t F_{ViT}^{(k)}) + F_{CNN}^{(l)} \quad (10)$$

Here, W_t maps Transformer embeddings into CNN channel dimensions, ψ reshapes token embeddings into spatial feature maps and $\tilde{F}_{CNN}^{(l)}$ denotes CNN feature maps enriched with Transformer-derived global context.

This feedback loop ensures that CNN activations, which are inherently local, are refined by lesion-wide structural context.

3.6 Final Fusion Representation

The fused representation, denoted in the equation (11), results from the concatenation of CNN-based and ViT-based enriched embeddings:

$$F_{fusion} = \tilde{F}_{CNN}^{(l)} \oplus \tilde{F}_{ViT}^{(k)} \quad (11)$$

where $\oplus$ denotes concatenation. Batch normalization and dropout ($p = 0.3$) are applied to stabilize learning. F_{fusion} thus contains both local lesion structures and global context, offering a holistic representation for classification.

Cross-level fusion provides advantages beyond simple late fusion or feature concatenation. It facilitates real-time interaction between local and global feature extractors, ensuring that:

- Transformers do not ignore subtle lesion details (like streaks or dots).
- CNNs are not blind to large-scale lesion symmetry or irregular spread.

This mimics the diagnostic reasoning process of dermatologists, who simultaneously focus on lesion details and its overall geometry.

3.7 Unified Feature Embedding and Classification

The unified feature vector F_{final} is passed into a classification head consisting of fully connected layers with dropout and batch normalization. Output of final dense layer is expected to be a probability distribution over the C diagnostic categories using softmax, as in (12)

$$\hat{y} = \text{softmax}(W_c F_{final} + b_c) \quad (12)$$

Where, W_c and b_c are classifier parameters, $\hat{y} \in \mathbb{R}^C$ is the predicted probability distribution.

To address dataset imbalance, particularly the under-representation of melanomas in HAM10000, a class-weighted categorical cross-entropy loss (13) is employed:

$$L = - \sum_{i=1}^C w_i y_i \log(\hat{y}_i) \quad (13)$$

Where, y_i = ground truth label for class i , $\hat{y}_i$ = predicted probability and w_i = class-specific weight inversely proportional to its frequency.

This ensures that rare but clinically critical categories, such as melanoma, receive higher importance during training

3.8 Training Configuration

The GAN network used here was trained using the Adam optimizer with a learning rate of 2×10^{-4} , $\beta_1 = 0.5$, and $\beta_2 = 0.999$. The batch size and epochs were selected as 16 and 100 respectively and the reconstruction-loss weight was set to $\lambda = 100$.

The proposed hybrid model was trained in two sequential steps. In the first step distortions were introduced on images from the training data subset of HAM10000 dataset and the resulting paired clean-distorted images were used to train the GAN. The trained GAN was used to pre-process the input images, in the second step. The pre-processed images from GAN will be then supplied to the hybrid ConvNeXtV2-Tiny and ViT-B/16 branches, followed by cross-level feature fusion and classification.

Key hyperparameters used for training the classification algorithm are:

- Optimizer: AdamW, chosen for its ability to decouple weight decay from gradient updates.
- Learning Rate: initialized at 3×10^{-4} with cosine annealing decay.
- Batch Size: 32, balancing GPU memory constraints with gradient stability.
- Epochs: 100.
- Dropout: applied at multiple stages (0.3 in CNN fusion, 0.5 in classifier) to reduce overfitting.
- Weight Decay: 1×10^{-2} , preventing over-parameterization.

PyTorch served as the primary framework for building and training the model. Training is conducted on Intel Ultra 9 Processor with GPUs with 32 GB memory with same environment for both GAN training and the hybrid CNN-ViT model training.

.

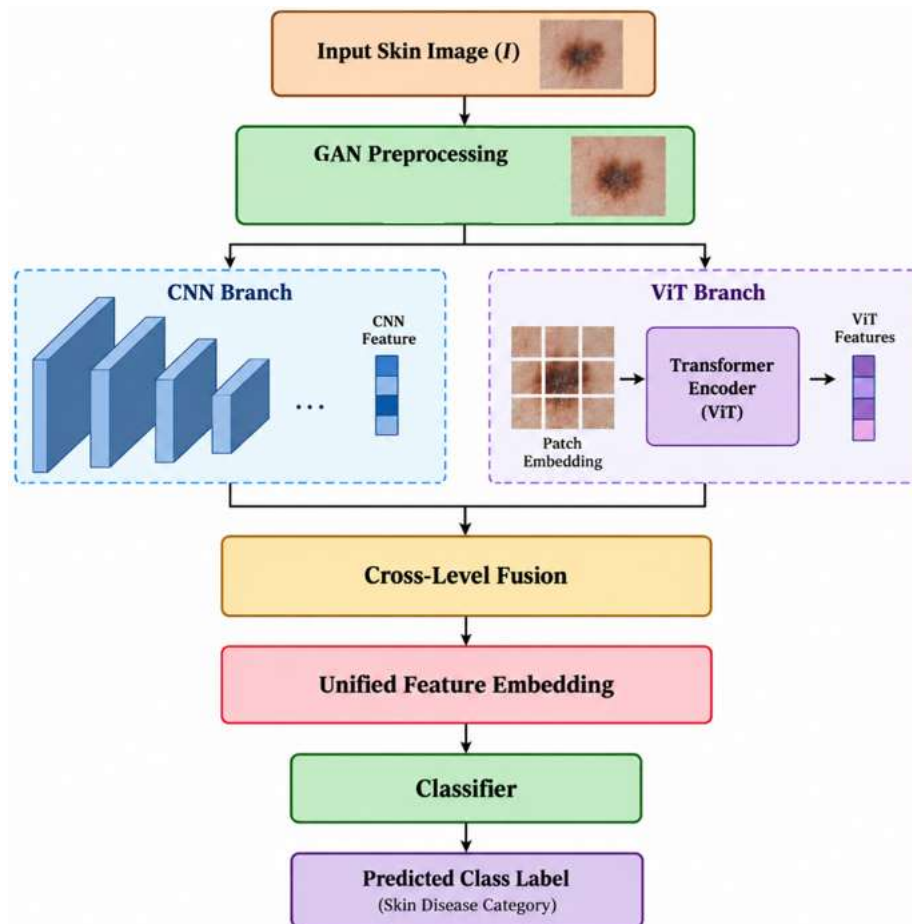


Figure 1: Proposed methodology

4. Results and Discussion

The proposed system was systematically evaluated using the HAM10000 dataset. To evaluate the robustness and the classification performance on variation in image quality, the assessment considered two conditions: (i) clean test sets, where images were presented without artificial distortions, and (ii) distorted test sets, where Gaussian noise, motion blur, and illumination variation were introduced.

4.1 Construction of Distorted Test Sets

To rigorously evaluate robustness, distorted test sets were generated from the original clean HAM10000 test dataset images by introducing controlled perturbations. The distortions included:

- **Gaussian Noise:** zero-mean Gaussian noise with variance $\sigma^2 = 0.01$, simulating sensor noise and acquisition imperfections.
- **Motion Blur:** linear motion blur with a kernel length of 7 pixels applied at random orientations, simulating hand tremor or camera shake during image capture.
- **Illumination Variation:** brightness and contrast shifts modeled as $I' = \alpha I + \beta$, where $\alpha \in [0.8, 1.2]$ adjusts contrast and $\beta \in [-20, 20]$ alters brightness, reflecting variable lighting conditions in clinical practice.

The same number of samples was kept in the distorted dataset similar to the clean dataset to make sure that there was one-to-one correspondence between samples. This enabled a direct comparison of the performance of the model under clean and distorted conditions.

For testing the GAN which was already trained, both the clean and distorted test image were put through the it. Enhanced images from the output of the GAN were then supplied to the CNN and ViT branches followed by cross-level fusion and classification. Representative images from both the clean and distorted test sets are shown in Figure 2.

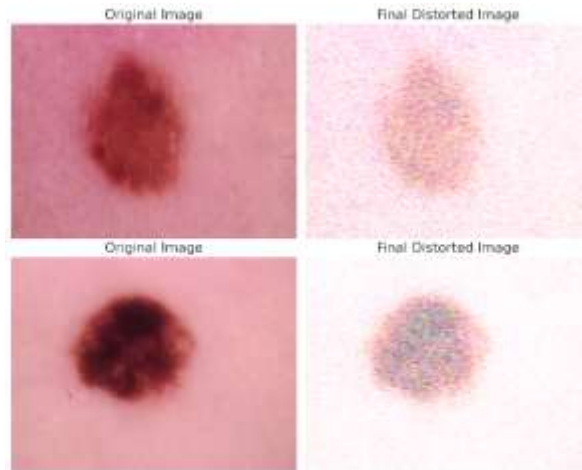


Figure 2: Original vs. Distorted Images

4.2 Evaluation Metrics

Several quantitative measures were used to provide a comprehensive review of results. Global performance indicator was accuracy and precision was defined as reliability of positive predictions. Recall was used to indicate the effectiveness of the model in identifying the true disease cases and the F1-score was used to balance between precision and recall. Other than this, the area under the receiver operating characteristic curve (AUC) yielded a measure of discrimination over thresholds.

4.3 Performance of the Proposed Model

The proposed model generates various outcomes, which are summarized in Table 1, the outcomes of combining GAN preprocessing for image enhancement and cross-level fusion between CNN and ViT backbones.

Table 1. Performance of the proposed model on HAM10000 dataset

Dataset	Condition	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
HAM10000	Clean	98.12	98.30	97.90	98.10	0.989
HAM10000	Distorted	94.60	95.00	94.20	94.55	0.971

The parameters shown clearly show the power of the proposed methodology. The model performance was 98.12% accurate when the data was clean, and the F1 score was 98.10% on the HAM10000 dataset with an AUC of 0.989. The combined pipeline achieves values higher than the majority of CNN and ViT baselines tested in this study, demonstrating that it provides a good trade-off between precision and recall. However, when distortion was introduced, the level of accuracy was reduced only slightly to 94.60%, demonstrating high resilience.

The performance results achieve show that the proposed framework achieves high accuracy of classification even in the presence of distortions in the input images.

4.4 Baseline Comparisons

A set of baseline experimental studies was used to analyse the impact of proposed methodology. In these baseline experiments we offer a comparison of the proposed approach with CNN-only, ViT-only and traditional CNN-ViT fusion methods. The accuracy and other parameter values are obtained as shown in figure 3 for different algorithms considered for clean and distorted data.

Table 2. Assessment of Results

Model Variant	Condition	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
CNN-only (ConvNeXtV2)	Clean	89.6	90.2	88.7	89.4	0.932
	Distorted	83.1	83.5	82.8	83.1	0.912

ViT-only (ViT-B/16)	Clean	90.4	91.0	90.0	90.5	0.938
	Distorted	82.7	83.1	82.2	82.6	0.914
CNN-ViT Fusion (no GAN)	Clean	92.1	92.5	91.7	92.1	0.955
	Distorted	87.3	87.9	87.0	87.4	0.932
Proposed method	Clean	98.12	98.30	97.90	98.10	0.989
	Distorted	94.60	95.00	94.20	94.55	0.971

The baseline experiments reveal crucial insights.

1. CNN-only vs ViT-only.

ConvNeXtV2 (CNN-only) performed well on clean images, with an accuracy of 89.6%, but it was far less successful in the distortion category (−6.5%). This is predictable given the fact that CNNs are effective in local feature extraction, and are influenced by blur and noise that may mask fine detail in the image. The results on clean data (90.4% accuracy) were slightly better than the ViT-only baseline (ViT-B/16), while the drop under distortion was larger (−7.7%), showing that transformers are sensitive to corrupted inputs as heavily relying on tokenized inputs.

2. CNN-ViT Fusion without GAN.

The introduction of fusion resulted in a significant increase in clean performance, to 92.1% accuracy and with a drop in distortion induced percentage of −4.8%. This shows that combining the local robustness of CNNs with the global modeling of ViTs yields complementary robustness. But with the absence of GAN preprocessing, the distortions reduced the robustness as well, which indicates that input enhancement is needed for stable performance.

3. Proposed Model.

The full pipeline consistently outperformed all baselines by achieving an accuracy of 98.12% and 94.60% under clean and distorted conditions respectively, compared to the other methods, indicating the benefit of GAN-based preprocessing introduced here in the proposed framework.

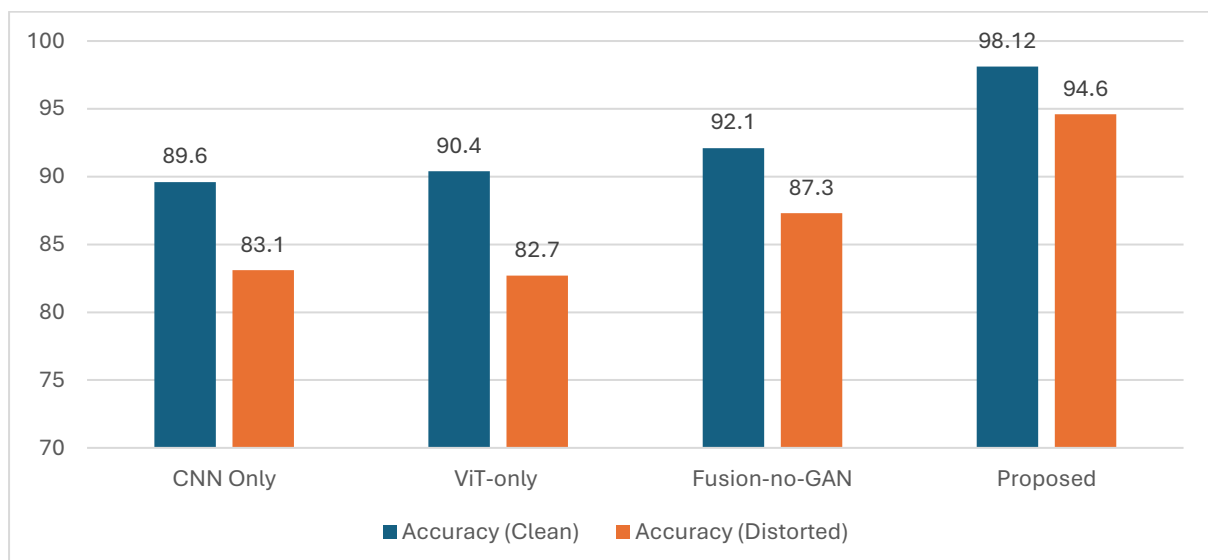


Figure 3: Clean vs Distorted Accuracy

4.5 Comparison with Literature

To evaluate how well the proposed model performs, its results were benchmarked against those documented in existing literature, with particular emphasis on research addressing skin lesion classification. Four key studies were chosen to balance historical significance with methodological relevance. These include CNN-based approaches, recent transformer-based models and hybrid CNN-Transformer fusions.

Table 3. Comparative Analysis with Related Literature.

Reference	Model / Approach	Dataset	Accuracy (%)
Srinivasu et al.[20]	MobileNetV2 + LSTM	HAM10000	85.34
Al-Masni et al.[21]	FrCN+DenseNet201	ISIC	88.70
Agyenta et al.[22]	InceptionV3 + ResNet50 + DenseNet201	HAM10000	86.91
A.Ali et.al.[23]	CNN+ViT	HAM10000	96.74
Proposed (this study)	GAN + CNN-ViT Cross-Level Fusion	HAM10000	98.12

From Table 3, we can infer the following.

The proposed model achieved an accuracy of 98.12% outperforms different prior studies reported in Table 3. This improvement can be attributed to the combined use of GAN preprocessing, cross-level CNN-ViT feature fusion.

4.6 External Cross-Dataset Validation

To assess the generalizability of the proposed framework an independent external cross-dataset validation is carried out using ISIC dataset, which is publicly available. Images corresponding to the image classes available in the HAM10000 dataset were only used for this external cross-dataset validation. These images were then supplied to the already trained GAN and the hybrid classification model. The proposed hybrid model achieved an accuracy of 96.20%, precision of 97.10%, F1-score of 96.30%, recall of 95.60% and AUC of 0.981. These results clearly indicate that the proposed model can perform well on independent datasets other than the HAM10000 dataset, therefore the model has a capability to generalize to an external dataset .

Conclusion

This study presented a hybrid distortion-robust skin disease classification framework that integrates GAN-based preprocessing and cross-level CNN-Vision Transformer (ViT) feature fusion. The methodology was designed to address two critical gaps in existing literature: (i) the vulnerability of CNN and ViT models to distortions and imaging artifacts and (ii) the lack of integration between local texture descriptors and global lesion context.

Extensive evaluations on HAM10000 dataset under both clean and distorted test conditions demonstrated that the proposed model consistently outperformed CNN-only, ViT-only, and naive fusion baselines. The proposed model produced an accuracy of 98.12% and F1-score of 98.10% on the clean dataset, and an accuracy of 94.60% and F1-score of 94.55% on distorted dataset

External cross-dataset validation of proposed model was also carried out on ISIC dataset, which produced an accuracy of 96.20% and F1-score of 96.30%, indicating the strong performance of the proposed model on an independent dataset. The combination of GAN-based preprocessing and cross-level CNN-ViT fusion contributes to improved robustness and generalization.

Overall, the proposed framework demonstrated a promising approach for computer-aided skin disease detection from dermoscopic images. Although the framework demonstrates strong performance on benchmark datasets, its clinical usability has not yet been established. The model should therefore be considered a research-oriented decision-support approach rather than a clinically validated diagnostic tool. Further prospective clinical validation using diverse patient populations and real-world clinical data is required before clinical deployment.

Data availability statement

The datasets utilized in this study are publicly available. The HAM10000 dataset can be accessed at <https://www.kaggle.com/kmader/skin-cancer-mnist-ham10000>, and the ISIC dataset is available at <https://challenge.isic-archive.com/data/>

References

- [1] Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017. doi: 10.1038/nature21056, 2017

- [2] K. Shehzad, T. Zhenhua, S. Shoukat, A. Saeed, and I. Ahmad, "A deep-ensemble-learning-based approach for skin cancer diagnosis," *Electronics*, vol. 12, no. 6, p. 1342, 2023. <https://doi.org/10.3390/electronics12061342>
- [3] N. Codella et al., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 ISIC workshop," 2018 IEEE 15th ISBI, pp. 168–172, 2018. doi: 10.1109/ISBI.2018.8363547
- [4] N. Gessert, T. Sentker, "Skin lesion classification using CNNs with patch-based attention," *IEEE Transactions on Biomedical Engineering*, 67(2):495-503, February 2020. <https://doi.org/10.1109/tbme.2019.2915839>
- [5] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint, <https://doi.org/10.48550/arXiv.2010.11929>
- [6] G M S. Himel et al., "Skin Cancer Segmentation and Classification Using Vision Transformer for Automatic Analysis in Dermatoscopy-Based Noninvasive Digital System," *Int J Biomed Imaging*. 2024 Feb 3; 2024:3022192. doi: 10.1155/2024/3022192
- [7] F S. Khan et al., "Transformers in medical imaging: A survey," *Medical Image Analysis*, vol. 88, p. 102802, 2023. <https://doi.org/10.1016/j.media.2023.102802>
- [8] A. Mahbod et al., "Fusing fine-tuned features for skin lesion classification," *Computerized Medical Imaging and Graphics*, vol. 71, no. 11, p. 2203, 2021. <https://doi.org/10.1016/j.compmedimag.2018.10.007>
- [9] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv:2102.04306, 2021. <https://arxiv.org/abs/2102.04306>
- [10] Z. Tu et al., "MaxViT: Multi-axis vision transformer," arXiv preprint, arXiv:2204.01697, 2022. <https://doi.org/10.48550/arXiv.2204.01697>
- [11] Z. Liu et al., "A convnet for the 2020s," arXiv preprint, arXiv:2201.03545, 2022. <https://doi.org/10.48550/arXiv.2201.03545>
- [12] S. Woo et al., "ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders," arXiv:2301.00808, 2023. <https://doi.org/10.48550/arXiv.2301.00808>
- [13] R. Abbas et al., "Automated hair removal in dermoscopy images using shallow and deep learning neural architectures," *Computer Methods and Programs in Biomedicine Update*, Volume 4, 2023, 100109. <https://doi.org/10.1016/j.cmpbup.2023.100109>
- [14] P. Celebi et al., "A methodological approach to the classification of dermoscopy images," *Computerized Medical Imaging and Graphics*, vol. 31, no. 6, pp. 362–373, 2007. doi: 10.1016/j.compmedimag.2007.01.003
- [15] A. E. Voronin et al., "Image Segmentation of Skin Neoplasms Using the Active Contour Method," *Method. Phys. Atom. Nuclei* 85, 1956–1960 (2022). <https://doi.org/10.1134/S1063778822090411>
- [16] S M. Yasir et al., "Attention-Based Deep Feature Aggregation Network for Skin Lesion Classification," *MDPI Electronics*, 14(12), 2025. <https://doi.org/10.3390/electronics14122364>
- [17] S. Al-Masni et al., "Skin lesion segmentation in dermoscopy images with ensemble deep learning methods," *Expert Syst. Appl.*, vol. 113, pp. 63–78, Dec. 2018. <https://doi.org/10.1016/j.cmpb.2018.05.027>
- [18] P.R., Bipin, Anoop V., Ramu R., Santhi K., Upendra Kumar, and Sai Kiran Oruganti. 2026. "A Hybrid ConvNeXtV2–MaxViT Framework With CNN-Based Feature Refinement for Skin Lesion Classification". *Journal of Innovative Image Processing* 8 (3): 747-64. <https://doi.org/10.36548/jiip.2026.3.001>.
- [19] G. Halladja et al., "EA-GAN: Enhanced attention generative adversarial network applied to hair removal in dermoscopy images," *Signal, Image and Video Processing*, 268 (2025). <https://doi.org/10.1007/s11760-025-03816-2>
- [20] Srinivasu PN, SivaSai JG, Ijaz MF, Bhoi AK, Kim W, Kang JJ. Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM. *Sensors*. 2021;21(8):2852. doi:10.3390/s21082852.
- [21] Al-Masni MA, Kim DH, Kim TS. Multiple skin lesions diagnostics *via* integrated deep convolutional networks for segmentation and classification. *Comput Methods Programs Biomed*. 2020;190(1):105351. doi:10.1016/j.cmpb.2020.105351.
- [22] Agyenta C, Akanzawon M. Skin lesion classification based on convolutional neural network. *J Appl Sci Technol Trends*. 2022;3(1):21–6. doi:10.38094/jastt301121.
- [23] Ali, A., Shahbaz, H., Damaševičius, R. (2025). xCViT: Improved Vision Transformer Network with Fusion of CNN and Xception for Skin Disease Recognition with Explainable AI. *Computers, Materials & Continua*, 83(1), 1367–1398. <https://doi.org/10.32604/cmc.2025.059301>