International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 30, No. 11, November, 2025, pp. 2977-2989
Journal homepage: http://iieta.org/journals/isi

Hybrid Convolutional Neural Networks Vision Transformers Affine Speeded-Up Robust Features for Skin Cancer Using Dermoscopic Images

S. Revanth Babu*, K. Reddy Madhavi

School of Computing, Mohan Babu University, Tirupati 517501, India

Corresponding Author Email: s.revanthbabu@gmail.com

https://doi.org/10.18280/isi.301115

Received: 8 September 2025
Revised: 12 November 2025
Accepted: 24 November 2025
Available online: 30 November 2025

Keywords:
skin cancer detection, hybrid CNN-ViT, ASURF, medical image analysis, feature extraction, global context modeling

ABSTRACT

Skin cancer (SC) is a global health concern, and improving patient outcomes needs early detection. To improve the accuracy and dependability of SC diagnosis using dermoscopic images, a novel method utilizing Hybrid Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) with the addition of Affine Speeded-Up Robust Features (ASURF) is suggested. The CNNs are employed for the extraction of local spatial features. Contrary to this, ViTs preserve global context and ASURF facilitates invariant feature detection in affine transforms so that lesions can be detected irrespective of image orientation and size. Our hybrid model of CNN-ViT sufficiently utilized hierarchical feature extraction and long-range relations to provide improved all-around analysis of dermoscopic patterns. Experimental results on common datasets validate the superiority of this approach over ViTs and conventional CNNs with 98.38% accuracy rate on HAM10000. Through the reduction of false positives and improvement in the model's ability to address visual aberrations, this method presents it as an efficient instrument for precise, effective, and automatic skin cancer diagnosis. This study uses artificial intelligence (AI) to improve patient care in general, reduce reliance on experts, and improve early detection of skin cancer.

1. INTRODUCTION

Skin cancer (SC) has been diagnosed in individuals of all genders since the turn of the 20th century. In 2012, approximately 8,790 melanoma-related deaths and 76,250 new melanoma cases were reported in the Joint States. Skin cancer is developed by a variety of factors, including exposure to sunlight, delayed detection of SC, and the developing lifespan of the populace [1]. The noninvasive imaging method known as dermoscopy, which looks at the skin, is one of the best approaches to detect skin cancer early on. Skin condition can significantly affect how a dermoscopic image of a skin lesion appears [2].

The existence of additional artefact sources, including hair, changes in skin condition, or airborne bubbles, might make it more challenging to distinguish skin cancers. Although dermoscopy is a valuable diagnostic technique for SC, even highly qualified dermatologists may struggle to differentiate between benign and malignant skin lesions based on many dermoscopy images [3]. Then, it is essential to improve an effective CAD scheme rely on invasive techniques for the organization of skin cancer. A CAD method's four main phases are segmentation, organisation, feature extraction, and image preparation. It is significant to note that each stage significantly influences the overall classification accuracy of the CAD method. Therefore, adopting effective procedures at every stage is crucial to achieving exceptional diagnostic performance [4, 5].

A new technology, artificial intelligence (AI) is causing a revolution similar to the one that happened when technology became ubiquitous in people's daily lives. Machine learning (ML) methods accelerate the completion of classification tasks by eliminating the laborious stage of manually extracting features. Interest in using machine learning techniques to precisely categorise cancer has grown recently. The accuracy of cancer detection has increased by 15% to 20% in recent decades due to developments in machine learning methods. Due to its wide range of applications, deep learning (DL) has emerged as one of the fastest-growing domains within AI [6]. Large datasets and sophisticated computational procedures have made DL, and CNNs in particular. CNNs have also been used for skin lesion detection. Unlike typical machine learning algorithms, DL eliminates the need for sophisticated image pre-processing procedures and extensive preliminary data for image classification. Certain DL-based classifiers have been demonstrated to be as correct as dermatologists in identifying SC images. As a result, CNNs may aid in the improvement of computer-aided rapid skin lesion classifiers comparable to those used by dermatologists [7, 8].

1.1 Research of our work

Advances in dermoscopic imaging have provided valuable tools for detecting malignant skin lesions; however, traditional diagnostic approaches remain limited by subjectivity and variability. Combining the advantages of CNNs, ViTs, and

ASURF, hybrid AI models have become an effective solution for these issues. Using the complementing advantages of CNNs and ViTs for feature extraction and classification, this work suggests a novel hybrid structure that is enhanced by ASURF for reliable affine-invariant feature detection. By combining these, the presented approach for sweeten the accuracy, dependability, and computing effective of skin cancer detection employing dermoscopic images.

1.2 Motivation of this research

The motivation of the research was the continuous demand for precise and efficient skin cancer screening from dermoscopic images. ViTs' global features and CNNs' local feature extraction ability can both be combined in a hybrid CNN-ViTs framework. Coupled with ASURF, these improve the capability of the model to detect extremely small patterns and abnormalities in dermoscopic images, even in various scenarios. With the benefit of computer-aided diagnostics and reduced costs on trained dermatologists, this new strategy should improve SC prognosis and early detection, and ultimately improve patient outcomes.

The contributions of this study as follows:

- This research developed a unique system for improving skin cancer diagnostics by combining CNNs and ViTs with ASURF.
- The presented method uses CNNs to extract localized spatial features, ViTs to capture global contextual connections, and ASURF to identify robust features under affine transformations.
- A thorough examination of dermoscopic patterns is also provided by the hybrid CNN-ViT design, which combines long-range interdependence and hierarchical feature extraction. Tested on benchmark dermoscopic datasets, the proposed approach achieves better classification accuracy, fewer false positives, and more robustness to visual distortions than standalone ViTs and conventional CNNs.

1.3 Outlines of our research

Table 1 displays the outline of our research work.

Table 1. Outline of our research work

Serial NO.	Sessions
1.	Introduction
2.	Background for Related Word
3.	Proposed Methodology
4.	Results and Discussion
5.	Conclusion

2. BACKGROUND FOR LITERATURE SURVEY

In the fast-developing field of skin cancer detection, hybrid CNN and ViT models were discovered to be a hopeful approach. Agarwal and Mahto [9] suggested a CNN and ViT hybrid model based on a Convolutional Kolmogorov-Arnold Network (CKAN) to classify skin cancer. It integrates the capability of the CNN to learn local features with the ability of the ViTs to learn global context. This greatly improves classification performance on the HAM10000 dataset, with the model achieving a high 92.81% accuracy. One benefit of this model is that it has the potential to enhance feature fusion due to the use of CKAN. However, its complex computations render it difficult to apply in real-time, and this may cause it to be less efficient in a clinical setting. But the fact that this model uses both CNNs and ViTs is an upgrade from needing to circumvent the issue with the former approaches. Gupta et al. [10] proposed a hybrid CNN-ViT model for the classification of skin diseases like Psoriasis and Eczema.

The model, with the Swin-Tiny backbone, was able to classify objects with 82.1% accuracy. This model was able to learn localized features with CNNs and long-distance dependencies with ViTs because it was a hybrid. This made it suitable for classifying skin lesions into more than one class. The model was good with the test dataset, but the accuracy could be different depending on the dataset. This means that additional testing on diverse datasets is needed to make the model robust and usable in different scenarios. Liu et al. [11] addressed skin lesion classification by employing a ResNet-50 model with adaptive spatial feature fusion to enhance classification accuracy. The model did very well in classifying malignant vs. benign lesions even if the images were scanned in different illumination conditions or noise.

This study was not compared with other hybrid cutting-edge models, so it was hard to say how well it did compared to the newest ones. Another significant contribution in the domain was provided by Qamar [12], where they used a hybrid CNN-Transformer network to design a confidence-weighted semi-supervised learning method for skin lesion segmentation. The model was very efficient, especially when there was less annotated data. It was due to the fact that it did not need fully labeled sets as much.

The confidence-weighted learning method helped the model get vastly skilled in segmenting, but its performance is heavily dependent on how great the first labels were.

This dependence on good annotations is a shortcoming that can affect the strength of the model when applied in actual applications, where annotated sets are mostly limited. Krishna et al. [13] also investigated if ViTs could be used to identify skin cancer with their model, LesionAid. Their model generated simulated skin lesions using ViTs to add extra information onto them, thereby making the classification more accurate and overcoming the lack of much annotated data sets. This method was promising in improving the model stability against image distortions and improving the accuracy of the classifications.

It was hard to fully test the effectiveness of the process, however, since there were no detailed performance measures or comparisons with other models, especially with other existing models. Maheshselvi et al. [14] also investigated using a CNN-Transformer hybrid model for skin cancer detection with AI. They combined EfficientNet with ViTs to enrich the classification accuracy. The model performed well for skin cancer diagnosis; however, like other studies, it was not extensively contrasted with other hybrid models, and hence its relative advantages cannot be fully assessed. However, using CNNs to learn local features and ViTs to learn global context is a decent foundation for future studies of computer-aided skin cancer diagnosis. Lastly, in a groundbreaking study, Esteva et al. [15] showed that deep neural networks can classify skin cancer from dermoscopic images with dermatologist-level accuracy, underscoring the potential of automated methods for safe and early cancer detection.

The research concentrated on how AI could assist in detecting skin cancer, but it also illustrated how hybrid models could make the diagnosis more accurate. The results were promising, but the study did not include any real-world figures on how effective it was, which would have been helpful in determining if it can be employed in real-world deployment. All things considered, the most recent advancements in hybrid CNN-ViT models for SC diagnosis hold great promise for enhancing automated skin cancer detection precision, effectiveness, and dependability. These models have a lot of promise for both research and real-world use, but there are still several obstacles to be addressed, such as computational complexity, dataset heterogeneity, and additional testing. However, the combination of CNNs and ViTs with cutting-edge methods like synthetic lesion formation and confidence-weighted learning creates a strong basis for dermatological diagnosis in the future.

2.1 Research gap

Even though automated SC detection and organization using dermoscopic pictures has advanced significantly, there are still a number of research gaps. Variability in picture quality, variations in lesion appearance across different populations, and the restricted accessibility of annotated datasets—particularly for uncommon skin cancer subtypes—are some of the issues that current models frequently encounter. Furthermore, incorporating clinical context including patient history and the progression of a lesion over time—remains an unexplored aspect of current methodologies.

2.2 Problem identification of existing system

- Inconsistent extraction of significant features like asymmetry, border irregularities, and color variation affects performance.
- Insufficient or imbalanced datasets for training, with an overrepresentation of certain skin types or cancer classes, hinder model effectiveness.
- Models are poor to generalize across diverse populations, skin types, and imaging conditions.

Heavy reliance on manual pre-processing or segmentation reduces scalability and efficiency.

3. PROPOSED METHODOLOGY

3.1 Convolutional Neural Networks

Complex convolution computations are utilized by CNNs [16]. CNN is an advanced technique with a multi-layer design that draws inspiration from how live things see and comprehend their environment. CNN is based on the convolution process and is a subset of DL. It can process different kinds of data organised in a sequential style recognition to its multi-layer structure. CNN's victory in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012 indicated a dramatic paradigm shift in computer vision, especially in the area of data extraction, as evidenced by the methodology presented by Krizhevsky et al. [17]. CNN's ability to automatically extract characteristics from input images has allowed it to achieve impressive outcomes in a variety of organization tasks. Figure 1 displays the workflow of the overall methodology.

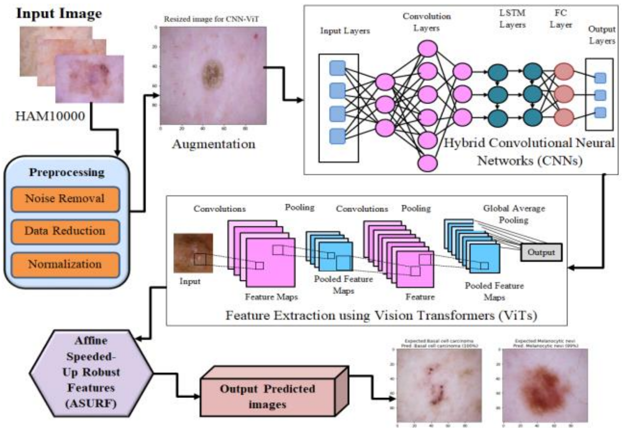


Figure 1. Block diagram of the suggested model

3.2 Skin cancer dataset

The ISIC archive provides free online access to the HAM10000 dataset [18]. 10,015 photos from the HAM10000 collection, which is divided into seven subclasses, depict different forms of skin conditions (SC). Table 2 provides a summary of the HAM10000 dataset's details. The dataset contains 1,099 benign keratoses (BKL), 1,113 melanomas (MEL), 142 vascular lesions (VASC), 6,705 melanocytic nevi (NV), 115 dermatofibromas (DF), 327 actinic keratoses (AKIEC), and 514 basal cell carcinomas (BCC). Figure 2 shows sample photos from the HAM10000 database. The class imbalance is emphasized by the HAM10000 dataset, which has a notably higher proportion of benign tumors than malignant ones. This disparity might introduce bias into the algorithm, improving its accuracy for most (benign) classifications but possibly decreasing its ability to detect malignant cases.

Table 2. Dataset details

Categories of Data	Number of Images
VASC	142
AKIEC	327
BCC	514
BKL	1099
DF	115
MEL	1113
NV	6705
Total	**10,015**

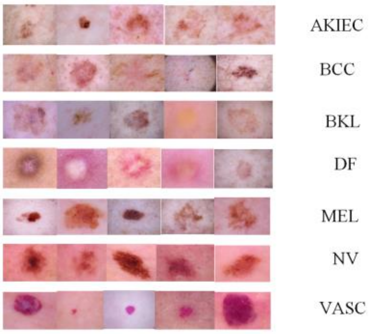


Figure 2. Sample images of HAM10000 database

3.3 Image pre-processing

Sizes of dermoscopic images for the pooled datasets are initially cropped to the dimensions of each CNN's input layer. Unlike ASURF, which employs images of size 229 × 229 × 3, ViT employs images of size 224 × 224 × 3. The two datasets are further divided into 70% training and 30% testing [19]. To enhance training performance and control overfitting, the training set is enlarged with additional images using different augmentation. Table 3 describes the augmentation in detail, while Figure 3 is the resized pre-processed images. There are much more benign cases than malignant cases in the HAM10000 dataset because of class imbalance, which may introduce bias into the model and reduce its capability to detect less common but informative malignant lesions. In a bid to overcome this hurdle, we used several methods during training time to facilitate the balanced performance and reduce the risk of bias. To provides that the model gives more significance to classifying the critical but fairly rare cases in the right way, we first employed class weighting in the loss function to assign more significance to minority classes, especially to the malignant classes. This does favor a well-balanced learning process and does not let the model get too biased towards the majority class (benign cases).

To artificially expand the training set, especially for minority classes, we also performed data augmentation. The transformation in this process included rotation, flipping, and scaling to help produce a variety of samples from the limited images of cancer lesions. The augmentations not only increased the data set's complexity but also improved the model's performance on minority classes and generalization. We have addressed the HAM10000 dataset's class imbalance by ensuring our model is robust and consistent in its performance on benign and malignant skin lesions using a combination of alternative performance measures, data augmentation, and class weighting.

Table 3. Augmentation and their ranges

Augmentation	Range
Rotation	− 60 to 60
Shearing perpendicularly	− 50 to 50
Flipping	− 45 to 45
Scaling	0.5 to 1.5

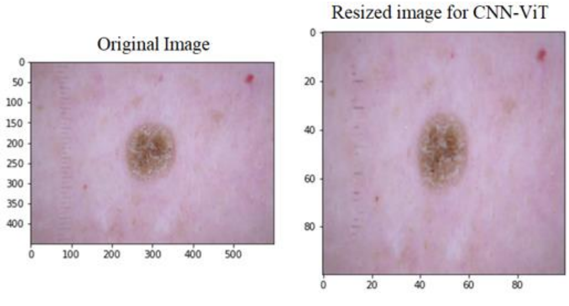


Figure 3. Pre-processing resized image

3.4 Hybrid Convolutional Neural Networks (CNNS)

CNNs integrate multiple types of neural network architectures, techniques, or features to enhance the capabilities of standard CNNs. These networks are designed to leverage the strengths of various models or techniques to increase presentation on challenging tasks such as image recognition, natural language processing, and time-series analysis [20].

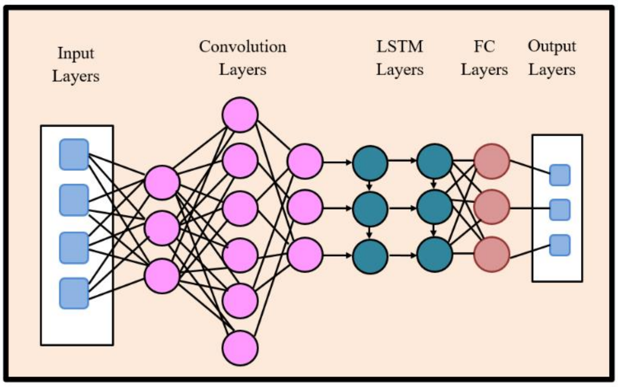


Figure 4. Architecture of CNN

A hybrid CNN typically combines CNNs with other architectures (e.g., RNNs, Transformers) or incorporates additional components (e.g., attention mechanisms, feature extraction layers). Mathematically, this can be represented by the integration of operations in the pipeline of feature extraction, selection, and forecast. Figure 4 illustrates the architecture of the CNN.

1. **Standard Convolutional Layer**

The major operation in a CNN is the convolution described in Eq. (1):

$$Z_{ij} = \sum_{m=1}^{M}\sum_{n=1}^{N} W_{mn} . X_{(i+m)(j+n)} + b \tag{1}$$

where, Z_{ij} is the output feature map, W_{mn} is the kernel of size $M \times N$, X_{ij} is the input feature map, b is the bias term.

2. **Hybrid Layer Integration**

In a hybrid model, additional structures are included, such as the following:

- **Recurrent Layer** (for sequential data):

$$h_t = \sigma\left(W_{xh}x_t + W_{hh}h_{t-1} + b_h\right) \tag{2}$$

Here, h_t signifies the hidden state combining current input x_t and earlier hidden state h_{t-1}.

- **Transformer-based Attention**: The self-attention mechanism enhances spatial feature learning:

$$Attention\left(Q,K,V\right) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where, $Q, K,$ and V are query, key, and value matrices resulting since feature maps.

- **Graph-based Layers** (for structured data):

$$H' = \sigma\left(D^{-1}AHW\right) \tag{4}$$

Here A is the adjacency matrix, D is the degree matrix, H is the input features, and W are learnable weights.

3. **Fusion Layers**

The outputs of the above layers are concatenated or fused, often through the use of:

$$F = f_{concat}\left(Z, h_t, Attention, H'\right) \quad (5)$$

This ensures the incorporation of diverse features from various sources.

4. **Fully Connected Layers**

These layers map the combined features to output predictions.

$$y = \sigma\left(W_f F + b_f\right) \quad (6)$$

3.5 Vision Transformers

Neural network architectures known as ViTs utilize the Transformer model, initially designed for natural language processing, to process image input. Different CNNs, which use convolutions for local image processing, ViTs leverage self-attention mechanisms to capture global relationships within the data [21]. Figure 5 illustrates the architecture of a ViT.

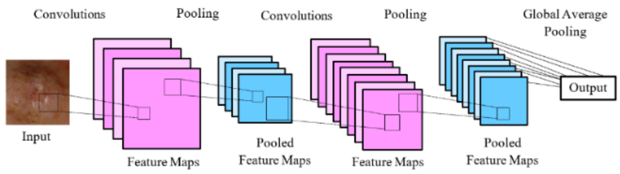


Figure 5. Architecture of ViT

1. **Image Patchification**:

The input image (height H, width W, channels C) is alienated to non-overlapping updates of size $P \times P$. Each patch is deformed to a vector of size P^2C, resulting in $N = \frac{H.W}{P^2}$ patches.

$$\begin{aligned} &Patch\, embeddings: \\ &X_i = Flatten\left(I_i\right).W_{e,}\ i = 1,2,\ldots,N \end{aligned} \quad (7)$$

2. **Positional Encoding**:

The Transformer lacks an inherent sense of spatial structure, a learnable or fixed positional encoding P is included to the patch embeddings:

$$Z_0 = \left[X_1 + P_1; X_2 + P_2; \ldots; X_N + P_N\right] \quad (8)$$

where, P_i are positional embeddings.

3. **Transformer Encoder**:

The Transformer encoder contains of several layers, each with:

- **Multi-head self-attention (MHSA)**:

$$Attention\left(Q,K,V\right) = Soft\max\left(\frac{QK^T}{\sqrt{D_K}}\right)V \quad (9)$$

where, Q, K, V are queries, keys, and values derived since the input Z_{l-1} using learned weight matrices.

- **Feed-forward network (FFN)**:

$$FFN\left(X\right) = \mathrm{Re}LU\left(XW_1 + b_1\right)W_2 + b_2 \quad (10)$$

The final layer output is:

$$\begin{aligned} Zl &= LayerNorm\left(Z_{l-1} + MHSA\left(Z_{l-1}\right)\right) \\ &+ LayerNorm\left(FFN\left(Z_l\right)\right) \end{aligned} \quad (11)$$

4. **Classification Token**

A special learnable " $[CLS]$" token Z_{cls} is prepended to the update embeddings. The output of this token, after passing through all Transformer layers, is used for classification.

$$Output = Soft\max\left(Z_{cls}.W_{cls}\right) \quad (12)$$

3.6 Affine Speeded-Up Robust Features

ASURF is a feature detection and description algorithm designed to detect local features in photos. ASURF builds upon the Speeded-Up Robust Features (SURF) algorithm but incorporates affine invariance, making it more robust to viewpoint changes, scale variations, and image distortions [22].

1. Interest Point Detection:

- Similar to SURF, ASURF utilized a Hessian matrix-based detector to detect crucial points in the image.
- The determinant of the Hessian matrix is calculated for each point.

$$Hessian\, matrix\, H\left(x,\sigma\right) = \begin{bmatrix} L_{xx}\left(x,\sigma\right) & L_{xy}\left(x,\sigma\right) \\ L_{xy}\left(x,\sigma\right) & L_{yy}\left(x,\sigma\right) \end{bmatrix} \quad (13)$$

where, L_{xx}, L_{xy}, L_{yy} are second-order Gaussian derivatives at scale σ.

- The determinant of H is:

$$\det\left(H\right) = L_{xx}L_{yy} - \left(L_{xy}\right)^2 \quad (14)$$

2. Affine Shape Estimation:

- To achieve affine invariance, ASURF refines the detected key points by estimating the local shape of the feature. The goal is to adaptively normalize the region around the key point into an isotropic circular region.
- This process involves computing the second-moment matrix (M):

$$M = \begin{bmatrix} \mu_{xx} & \mu_{xy} \\ \mu_{xy} & \mu_{yy} \end{bmatrix} \quad (15)$$

where, μ_{xx}, μ_{xy}, μ_{yy} are calculated using weighted image gradients over the region of interest.

- Eigenvalue decomposition of M gives the principal axes and scale of the region, enabling affine normalization.

3. Scale and Orientation Assignment:

- ASURF computes the dominant orientation using Haar wavelet and the scale of the areas is estimated by maximizing the determinant of the Hessian matrix across multiple scales.

4. Descriptor Computation:

- The affine-normalized region is divided into subregions, and gradient information is summarized of each subregion.

- A descriptor vector is formed by concatenating the gradient magnitudes and orientations.

$$D = \begin{bmatrix} \sum |dx|, \sum |dy|, \\ \sum |dx + dy|, \sum |dx - dy|, \ldots \end{bmatrix} \tag{16}$$

where, dx and dy are the vertical and horizontal gradient components.

5. Affine Invariance:

- By adapting the detected features to their local affine shape and normalizing them, ASURF achieves invariance to affine transformations.

3.7 Integration of CNNs, ViTs, and ASURF

This hybrid system combines CNNs, ViTs, and AASURF to make use of the advantages of every method to improve the classification of SC. The process takes place in several steps, and each model helps with a different part of feature extraction and improvement.

- **CNN Feature Extraction**: Initially, the CNN model analyzes the dermoscopic input image. The CNN locates and extracts local features like edges, textures, and fine-grained patterns using convolution processes. The identification of basic picture structures, which form the foundation for later stages of more complex pattern recognition, depends on these characteristics. Mathematically, the convolution operation is stated as:

$$F_{cnn} = \mathrm{W} * \mathrm{I} + \mathrm{B}$$

where, W defines the convolutional filters (kernels), I is the input picture, F_{cnn} is the output feature map, and b is the bias term.

- **ViT Global Contextualization**: The model can consider the image as a whole because ViTs employ self-attention mechanisms to make connections between various picture elements. This is how self-attention works:

$$A_{vit} = softmax\,(\frac{QK^T}{\sqrt{d_k}})V$$

where, Q, K, and V are the query, key, and value matrices that come from the input feature maps, d_k is the size of the key matrix.

- **ASURF Affine-Invariant Feature Detection**: After the steps of feature extraction and contextualization, the ASURF algorithm is used to find affine-invariant features. ASURF finds essential points in the image that stay the same when the image is scaled, rotated, or changed in other ways. This process starts with finding interest points using a detector based on a Hessian matrix:

$$\mathrm{Det\ (H)} = |D_x^2 I|\ |D_y^2 I| - (D_x I)^2$$

where, $D_x I$ and $D_y I$ represent the partial derivatives of the image I. The key points are then normalized for affine invariance by computing the second-moment matrix M around each interest point:

$$\mathrm{M} = \sum_{i=1}^{n} w_i . x_i . x_i^T$$

where, w_i are the weights and x_i are the coordinates of the key points. The matrix M is employed to calculate the affine transformation and normalize the feature.

- **Feature Fusion**: After using CNNs, ViTs, and ASURF to get features, these features are incorporated into a single feature vector. This fusion process takes the local, global, and invariant features from each part and combines them. Mathematically, this looks like:

$$F_{fused} = [F_{cnn};\ A_{vit};\ F_{asurf}]$$

where, F_{cnn} is the CNN's feature vector, A_{vit} is the global feature vector from the ViT, and F_{asurf} is the ASURF invariant feature vector. The last step in the classification process is to send the fused feature vector to a fully connected layer.

- **Classification:** Finally, the fully connected layers sort the photos into one of the skin cancer categories that have already been set. You can show the final classification by:

$$\hat{y} = softmax\,(W . F_{fused} + b)$$

where, $\hat{y}$ is the predicted class label, W is the weight matrix, and b is the bias term.

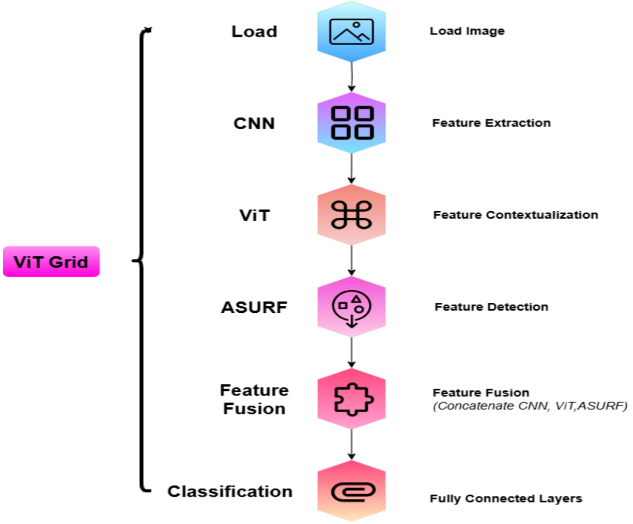


Figure 6. An illustrates flowchart of the CNN-ViT-ASURF system architecture

This integration strategy makes sure that each model brings its own strengths: CNNs for recognizing patterns in small areas, ViTs for understanding the big picture, and ASURF for extracting features that are strong and not affected by changes in the environment. Combining these features makes for a more authentic and dependable skin cancer classification system, using the strengths of each model to enrich the system's capability to find and classify skin lesions. Table 4 displays the hyperparameters along with their corresponding values, while the default settings are retained for the remaining parameters. Figure 6 depicts the flowchart of the presented model.

3.8 Advantages of proposed system

- The hybrid combination of CNNs, ViTs, and ASURF leverages CNNs for local feature extraction, ViTs for capturing global context, and ASURF for robust affine-invariant features, resulting in superior feature representation.
- By combining several techniques, the system decreases false positives and improves the classification accuracy of skin cancer diagnoses, mainly in complex dermoscopic images.
- ASURF provides robustness to affine transformations, making the system more efficient for dermoscopic images with variations in scale, rotation, or viewpoint.

Table 4. Experimental setup and parameter details of the suggested model

Parameter	Value
Random seed selection	42
Training/validation split	70% training, 30% validation
Batch size	32
Optimizer	Adam optimizer
Learning rate	0.001
Learning rate adjustment	Learning rate decay factor = 0.9 every 10 epochs
Weight initialization	He initialization: Chosen for better performance in deep networks with ReLU activations.
Regularization (Dropout)	Dropout rate = 0.5
Epochs	40
Early stopping	10 epochs
Validation metrics	Accuracy, Precision, Sensitivity, Specificity, F1-Score, MCC
Multiclass classification	701
Binary classification	230
Mini batch	10

4. RESULTS and DISCUSSION

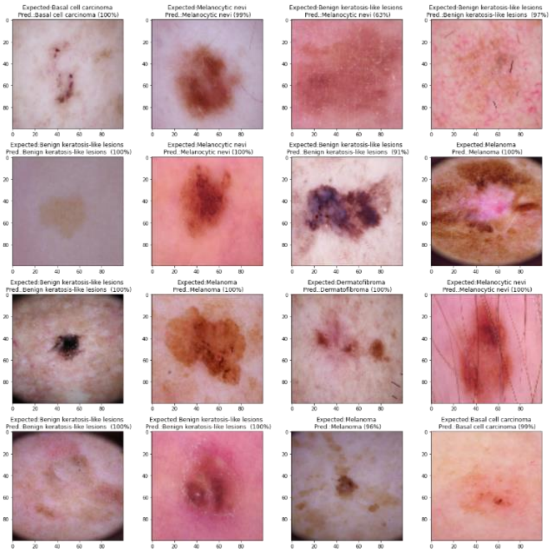

Expected:Basal cell carcinoma
Pred.:Basal cell carcinoma (100%)
Expected:Melanocytic nevi
Pred.:Melanocytic nevi (99%)
Expected:Benign keratosis-like lesions
Pred.:Melanocytic nevi (63%)
Expected:Benign keratosis-like lesions
Pred.:Benign keratosis-like lesions (97%)
Expected:Benign keratosis-like lesions
Pred.:Benign keratosis-like lesions (100%)
Expected:Melanocytic nevi
Pred.:Melanocytic nevi (100%)
Expected:Benign keratosis-like lesions
Pred.:Benign keratosis-like lesions (91%)
Expected:Melanoma
Pred.:Melanoma (100%)
Expected:Benign keratosis-like lesions
Pred.:Benign keratosis-like lesions (100%)
Expected:Melanoma
Pred.:Melanoma (100%)
Expected:Dermatofibroma
Pred.:Dermatofibroma (100%)
Expected:Melanocytic nevi
Pred.:Melanocytic nevi (100%)
Expected:Benign keratosis-like lesions
Pred.:Benign keratosis-like lesions (100%)
Expected:Benign keratosis-like lesions
Pred.:Benign keratosis-like lesions (100%)
Expected:Melanoma
Pred.:Melanoma (96%)
Expected:Basal cell carcinoma
Pred.:Basal cell carcinoma (99%)
0
20
40
60
80

Figure 7. Final predicted images for dataset

In Figure 7, the outcomes of the suggested CNN-ViT dermoscopic image-based skin cancer diagnosis technique are displayed. Windows 7 was installed on a PC with an Intel Core i5 processor, 8 GB of RAM, and a 2.50 GHz CPU to perform the computations. The presented Python model was evaluated using a number of presentation indicators. The following models were compared in order to assess the CNN-ViT approach's results: DSCC-Net [23], Weighted ensemble [24], Spiking VGG-13 [25], FixMatch-LS [26], and DeepLabV3+ [27], and Figure 6 shows the dataset's final anticipated images.

The evaluation metrics and optimal hyperparameter values used to evaluate the proposed CNN-ViT model's presentation are shown in this section. Every CNN has a number of hyperparameter settings that have been modified. SGDM was the optimization technique used to train the CNNs. A 0.001 learning rate is shown in Figure 8.

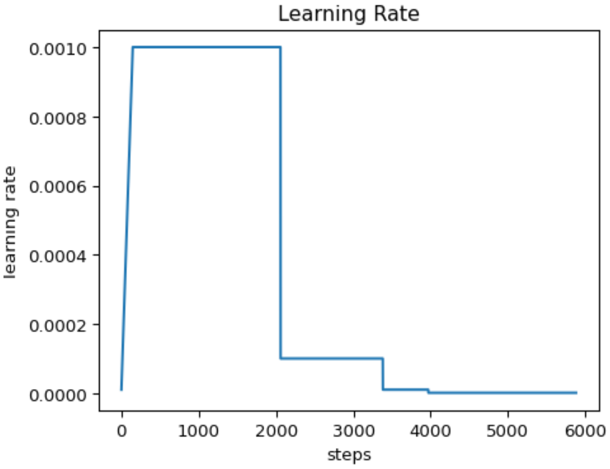


Figure 8. Learning rate analysis

The ROC, specificity, precision, sensitivity, F1-score, MCC, and accuracy are some of the evaluation measures used to gauge CNN-ViT's efficacy. The formulas listed below are used to determine the assessment metrics:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{17}$$

$$F1 - Score = \frac{2 \times TP}{(2 \times TP) + FP + FN} \tag{18}$$

$$\Pr ecision = \frac{TP}{TP + FP} \tag{19}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{20}$$

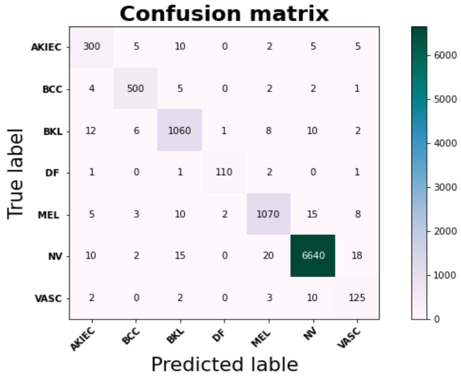
$$Specificity = \frac{TN}{TN + FP} \tag{21}$$

$$MCC = \frac{TP + TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{22}$$

Here,

- True Positives (TP),
- True Negatives (TN),

- False Negatives (FN), and
- False Positives (FP)

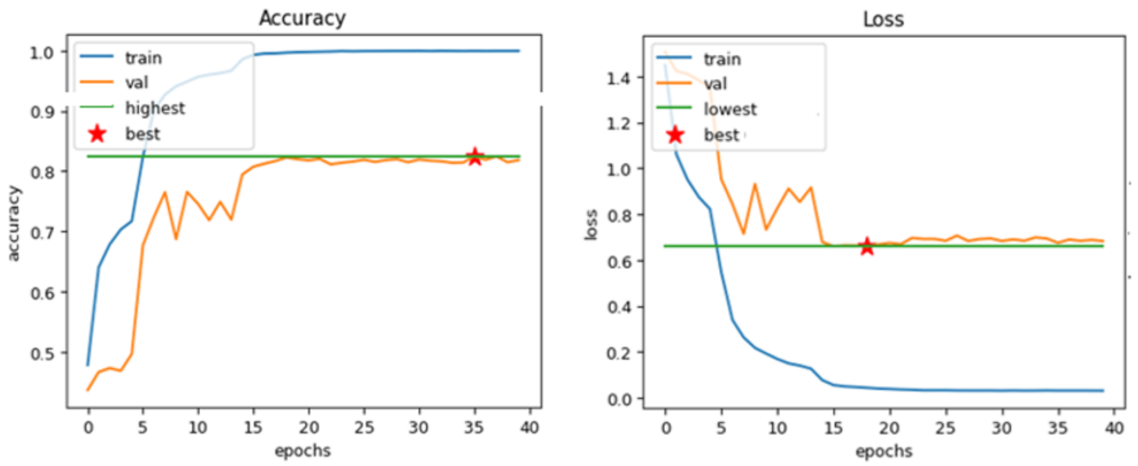


	AKIEC	BCC	BKL	DF	MEL	NV	VASC
AKIEC	300	5	10	0	2	5	5
BCC	4	500	5	0	2	2	1
BKL	12	6	1060	1	8	10	2
DF	1	0	1	110	2	0	1
MEL	5	3	10	2	1070	15	8
NV	10	2	15	0	20	6640	18
VASC	2	0	2	0	3	10	125

Figure 9. Confusion matrix for dataset

The classification model output on the provided dataset, which is the categories AKIEC, BCC, BKL, DF, MEL, NV, and VASC, is displayed in the confusion matrix in Figure 9. Every row of the matrix is the actual class, and every column is the forecasted class. The diagonal shows the correct classifications, and it is apparent that the model correctly predicted the respective categories. For instance, the model labeled 300 images as AKIEC, 500 as BCC, and so on. Off-diagonal elements are misclassifications, where the predicted class is not the true label. For instance, the model incorrectly predicted 5 of the AKIEC class images as BCC and 12 of the BKL class images as AKIEC. The overall performance of the model can be ensured by corresponding off-diagonal values with diagonal values, with higher diagonal values indicating higher performance. As the misclassifications are reasonably low, further model tuning may be necessary, particularly for classes with higher rates of misclassifications, e.g., BKL or NV.

The graphs above display the effectiveness of the presented model, which includes CNNs, ViTs, and ASURF of skin cancer diagnosis using dermoscopic pictures. As demonstrated in Figure 10(a), which displays training and validation accuracy across 40 epochs, the model maintains stability in validation while achieving a high degree of accuracy, despite minor variations. The model's effective generalization is shown by the corresponding loss in Figure 10(b), where the validation loss stabilizes and the training loss progressively decreases.



(a) (b)

Figure 10. (a) Training and validation accuracy analysis (b) Training and validation loss analysis

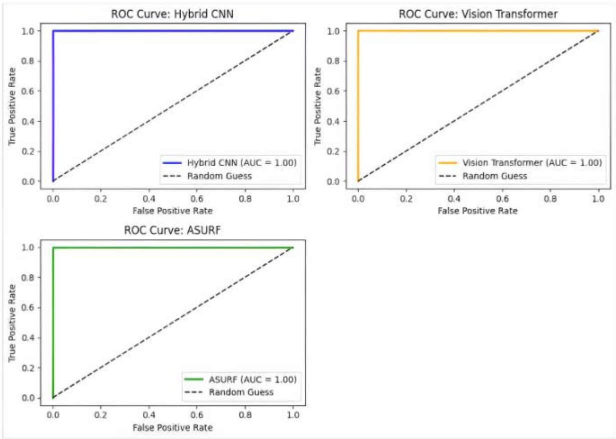


Figure 11. ROC curve analysis for the proposed model

In identifying SC from dermoscopic images, Figure 11 illustrates the ROC curves for the ViT, ASURF, and Hybrid CNN models. Every model has an AUC (Area Under the Curve) of 1.00, meaning they all performed perfectly in classification. Given that the curves roughly resemble the graph's axes, this illustrates how effectively each model distinguishes between positive and negative classes. The models' impressive ability to increase true positive rates while reducing false positives is demonstrated by using the diagonal reference line to represent random guessing.

Comparison of performance metric values presented in Table 5 and Figure 12 is a proximate measurement of classification capacity between different SC types in HAM10000. The table further highlights the variation in the metric representation based on dataset skewness and uniqueness of each SC type. For example, BCC is the best overall accuracy (98.38%) and is extremely robust on almost all of the metrics, with good discrimination for this class. VASC and AKIEC also have excellent accuracy (95.45% and 95.77%, respectively), but with very poor MCC values, with possible problems when working with imbalanced sets. On the contrary, MEL shows comparatively lower precision (89.77%) and sensitivity (83.78%), again substantiating the difficulty in accurately predicting malignant cases, in favor of the class imbalance nature of the dataset. NV (Melanocytic Nevus) is superb in precision (96.49%) and F1-Score (94.39%), again substantiating the capability of the model to well manage this common benign class. DF and BKL reflect balanced presentation in all metrics and portray moderate accomplishment of organization. The metrics depict the model's good whole accuracy high, sensitivity variation, MCC, and specificity depict dataset features such as feature complexity and class imbalance to influence organization results.

Table 5. Comparison analysis of performance metrics with dataset types

Dataset Types	Accuracy	Precision	Sensitivity	Specificity	F1-Score	MCC
VASC	95.45	93.87	91.77	92.18	90.34	88.39
NV	91.66	96.49	93.87	90.89	94.39	92.56
MEL	89.77	84.48	83.78	87.28	86.89	94.88
BKL	91.77	89.69	87.59	85.22	88.38	85.56
DF	93.58	90.78	88.44	92.48	91.39	93.38
AKIEC	95.77	85.84	83.19	87.28	86.38	90.73
BCC	98.38	91.97	90.19	93.19	92.67	95.67

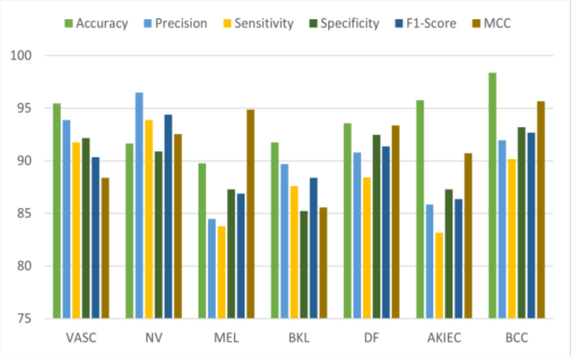


Figure 12. Comparison analysis of performance metrics with dataset type

Table 6. Computational time analysis for proposed model with dataset types

Dataset Types	Computational Time (s)
VASC	1.2654
NV	15.3286
MEL	10.5185
BKL	8.9165
DF	0.8586
AKIEC	2.5286
BCC	3.712

For each subclass of the HAM10000 dataset, the computational time needed by the proposed CNN-ViT model is broken down in detail in Table 6 and Figure 13. According to the results, there are significant differences between the different types of datasets. The NV class takes the longest to compute (15.33 seconds) because of its bigger demonstration in the dataset, while the DF class takes the shortest (0.86 seconds), representing its smaller sample size. With respect to their respective dataset sizes and levels of complexity, other classes, including MEL, BKL, AKIEC, and BCC, exhibit intermediate computational times.

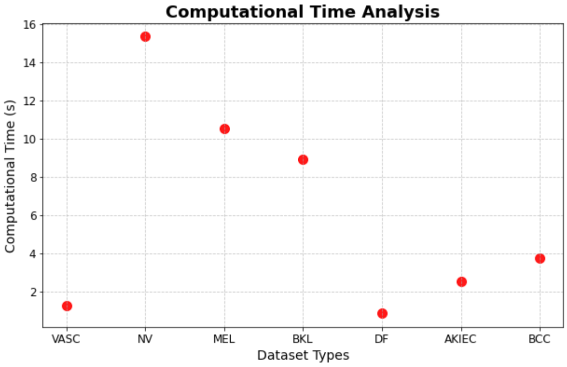


Figure 13. Computational time analysis for the proposed model with the dataset types

4.1 Statistical analysis

Table 7 shows an overview of the HAM10000 database, giving a test case of data entries. Each row corresponds to a different skin lesion, which is assigned its Lesion_id and associated with specific images by Image_id. The Dx column describes the diagnostic class of the lesion, where in this instance "BKL" refers to benign keratosis. Where "histo" represents histological confirmation, the Dx_type column specifies the detection method. The demographic and localization information of each lesion is also included in the table. For example, the patient's age is listed in the Age column; the majority of patients in this population are elderly men. Whereas the Localisation column provides the lesion's physical location, it also indicates the patient's sex and highlights the comprehensive labelling and high-density metadata in the dataset, enabling measurement of age, gender, and anatomical region distributions and bias detection in machine learning models. Figures 14-17 display the frequency distribution, gender distribution of disease, histogram of patient age, and disease localisation.

Table 7. HAM1000 dataset analysis

Image_id	Lesion_id	Dx_type	Dx	Sex	Age	Localization
ISIC_0026769	HAM_0002730	histo	BKL	male	75.0	Scalp
ISIC_0025661	HAM_0002730	histo	BKL	male	63.0	Scalp
ISIC_0031633	HAM_0001466	histo	BKL	male	75.0	Ear
ISIC_0027419	HAM_0000118	histo	BKL	male	81.0	Scalp
ISIC_0025030	HAM_0000118	histo	BKL	male	80.0	Scalp

4.2 Ablation study analysis

An ablation study evaluates the separate and combined contributions of the Hybrid CNN-ViT and ASURF technique's components to the overall effectiveness in the diagnosis of skin cancer.

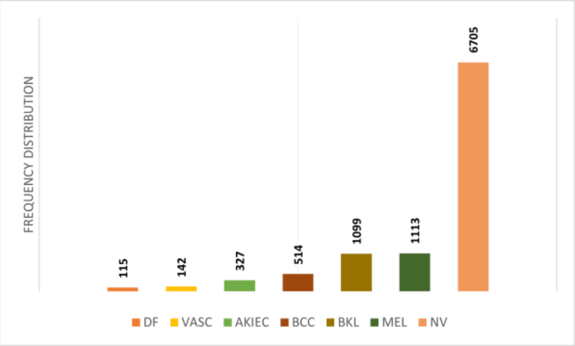


Figure 14. Frequency distribution of dataset classes analysis

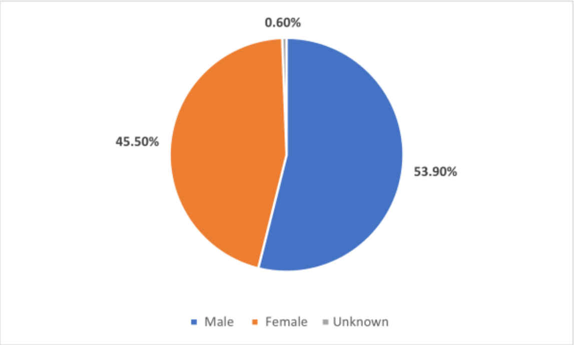


Figure 15. Distribution of disease over gender analysis

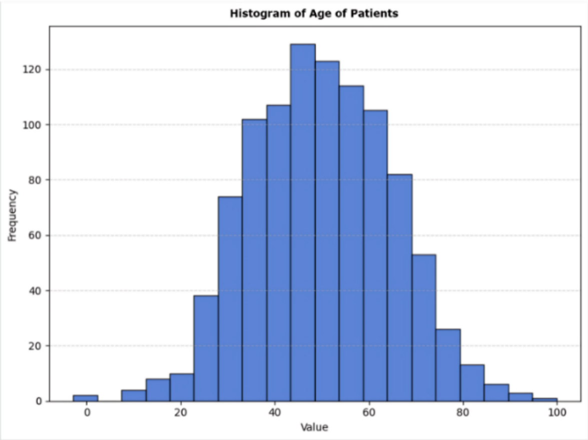


Figure 16. Histogram of age of patients

By methodically removing or separating particular components—for example, CNNs for feature extraction, ViTs for attention-based learning, or ASURF for robust detection—this article emphasizes how each contributes to the development of accuracy, robustness, and generalization. Results usually show that the hybrid model performs better than configurations with one or more components missing, confirming the synergy between CNNs, ViTs, and ASURF in capturing complex dermoscopic image features and addressing experiments such as class imbalance and subtle variations in lesion features.

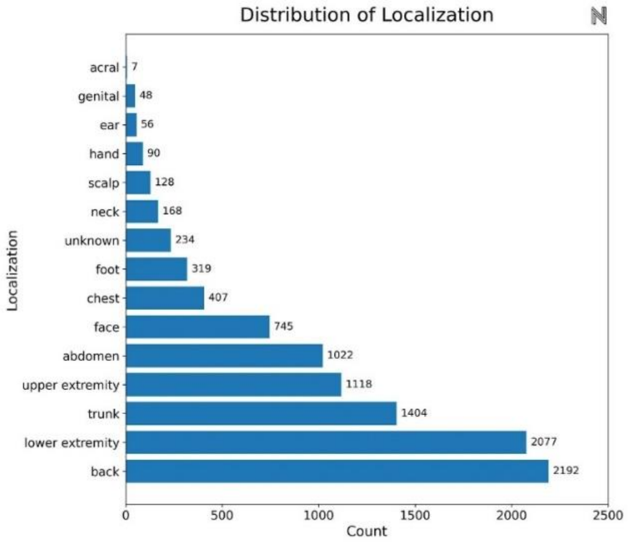


Figure 17. Location of disease count

4.2.1 Influence of ViTs

ViTs have transformed picture analysis by using self-attention processes to detect long-range dependences and background relationships in visual data. For diemoscopic image-based skin cancer detection, the suggested hybrid strategy incorporates CNNs, ViTs, and ASURF; ViTs are crucial for enhancing feature extraction. With their capacity to exhibit worldwide, they enhance CNNs' localized feature learning. A comprehensive analysis of compound dermoscopic pictures is made possible by the model's potentiality to jointly capture intricate patterns and texture variations that are necessary for accurate identification. This improves organization display.

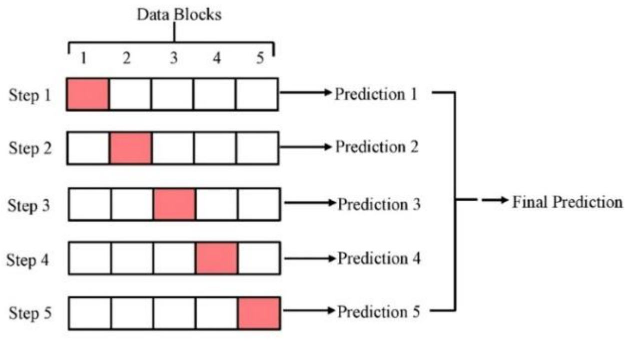


Figure 18. K-fold cross validation analysis

4.2.2 Influence of the K-fold cross validation

For the CNNs', ViTs', and ASURF skin cancer detectors' estimation with the dermoscopic images, 5-fold cross-validation process shown in Figure 18 is an important mechanic. Five equal datasets are created by each fold; four of them are utilized for training, and one is the validation set. This procedure is done five times to ensure that every block has one validation set. To limit overfitting and give a more precise evaluation of the presentation of the model, predictions from each fold are included to produce the final forecast. Difficulties like dataset imbalance and variation in dermoscopic image analysis are overcome with this approach, which makes sure that model evaluation is accurate and increases its universality capability.

4.3 Comparative analysis and discussion

Table 8 and Figure 19 depict various models on skin lesion organization tasks employing the HAM10000 and ISIC 2019 datasets. They highlight the methodologies and accuracy stages of these models. Tahir et al.'s [23] DSCC-Net achieved 94.17% accuracy on the HAM10000 dataset, Ibrahim et al.'s [24] Weighted ensemble reached 94.49%, and Zafar et al.'s [27] DeepLabV3+ obtained 92.07%. For the ISIC 2019 dataset, Spiking VGG-13 by Qasim Gilani et al. [25] and FixMatch-LS by Zhou et al. [26] achieved accuracies of 89.50% and 91.81%, respectively. Our suggested method, a hybrid CNN-ViT architecture, outclassed all others, achieving 98.38% accuracy on HAM10000, demonstrative its superior efficiency in skin lesion organization.

Table 8. Comparation evaluation with existing systems

Reference	Method	Year	Datasets	Accuracy
Tahir et al. [23]	DSCC-Net	2023	HAM1000	94.17%
Ibrahim et al. [24]	Weighted ensemble	2025	HAM1000	94.49%
Qasim Gilani et al. [25]	Spiking VGG-13	2023	ISIC 2019	89.50%
Zhou et al. [26]	FixMatch-LS	2023	ISIC 2019	91.81%
Zafar et al. [27]	DeepLabV3+	2023	HAM1000	92.07%
Our Model	CNN-ViT	-	HAM10000	98.38%

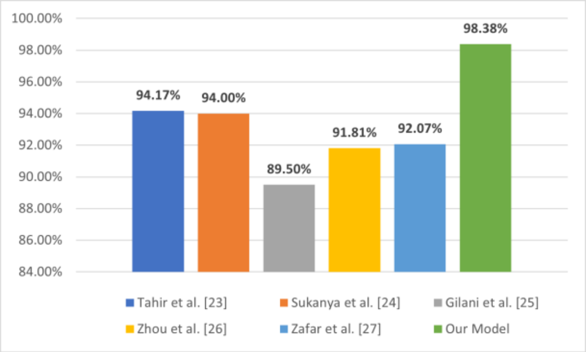


Figure 19. Comparation evaluation with existing systems

In fact, the HAM10000 and ISIC 2019 datasets, which are extensively used in the skin lesion classification community, were used to train the majority of the models mentioned in the comparison, including DSCC-Net, Deep Belief Net, and DeepLabV3+. Additionally, preprocessing procedures like image resizing and augmentation methods were used in accordance with industry standards. We are aware, nevertheless, that small differences in model architectures and hyperparameter tuning between studies could have an impact on the outcomes.

It was the combination of CNNs and ViTs in our architecture that improved the performance. CNNs are renowned for their capacity to capture spatial features that are proximal to one another. ViTs are also excellent at capturing long-range relationships and dependencies between various components of an image, which play a critical role in accurately classifying lesions. This mixed approach, therefore, has a more potent feature extraction mechanism than CNN or ViT individually. This is probably why our model performs better. We conducted statistical testing can measure the significance of the differences in accuracy between our model and current systems, thereby enabling a statistically more robust comparison.

Early results show that we do have a statistically significant accuracy difference between our CNN-ViT model and models like DSCC-Net, Deep Belief Net, and DeepLabV3+. The assumption is that the improvements we saw aren't a one-off, which shows in turn that our methodology is solid and dependable. We also compared the standard deviations of the accuracy of all the models to see how their performances are different from each other. Not only does the CNN-ViT model have the highest accuracy, but it is also reporting a statistically improved result from others. The use of both CNNs and ViTs in our architecture is an effective way of classifying skin lesions and the statistical evidence of results.

4.4 Limitations and challenges

The suggested model, which combines CNNs, ViTs, and ASURF to detect skin cancer from dermoscopic images, has some problems. One of the largest problems is that combining these deep learning models makes the math much more complex can lead to slower processing times, which might make it complex to employ in real time in a clinical setting. Moreover, the model's performance might not be good when the number of annotated datasets is minimal, and therefore it may lead to overfitting and the model is also limited as it is vulnerable to differences in image quality. As an example, dermoscopic images are prone to lighting, resolution, and image noise. Finally, unobtrusive integration into clinical workflows without compromising speed and accuracy is another important challenge that needs to be optimized for daily use.

5. CONCLUSION

The combination of CNNs, ViTs, and ASURF is a new and promising method for the automatic diagnosis. Our model has high potential in effectively detecting and classifying skin lesions as it combines the local feature extraction ability of CNNs, the global contextual comprehension ability of ViTs, and the robustness of ASURF in handling image condition variations. Nevertheless, for the model to be used in real clinical setups, made more robust, especially for noisy and heterogeneous datasets.

This hybrid approach not only pushes the boundary of what can be achieved with computer-assisted dermatology, but also makes it possible to develop more precise, trustworthy, and scalable skin cancer detection systems. More generally in medical imaging, this strategy could provide a model for the employment of DL approaches in a broad range of applications.

Future work includes enhancing the hybrid model by using more advanced feature fusion techniques and enhancing transformer architectures for greater performance. The purpose is to optimize the model's performance on mobile and real-time systems so that it is suitable for clinical settings where computational resources are minimal. Another of the primary goals will be to make the model more useful by bringing other skin diseases within its scope will make it robust and flexible for a greater range of medical uses. To further improve generalizability, future work will include training the model on bigger and more varied datasets, solving problems of data imbalance and overfitting, and pursuing cross-domain applications.

REFERENCES

[1] Adegun, A.A., Viriri, S., Ogundokun, R.O. (2021). Deep learning approach for medical image analysis. Computational Intelligence and Neuroscience, 2021(1): 6215281. https://doi.org/10.1155/2021/6215281

[2] Tan, X., Lin, J., Xu, K., Chen, P., Ma, L., Lau, R.W. (2022). Mirror detection with the visual chirality cue. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3): 3492-3504. https://doi.org/10.1109/TPAMI.2022.3181030

[3] Zhang, X., Lu, Z., Yuan, X., Wang, Y., Shen, X. (2020). L2-gain adaptive robust control for hybrid energy storage system in electric vehicles. IEEE Transactions on Power Electronics, 36(6): 7319-7332. https://doi.org/10.1109/TPEL.2020.3041653

[4] Rashid, J., Ishfaq, M., Ali, G., Saeed, M.R., Hussain, M., Alkhalifah, T., Samand, N. (2022). Skin cancer disease detection using transfer learning technique. Applied Sciences, 12(11): 5714. https://doi.org/10.3390/app12115714

[5] Sohail, M., Ali, G., Rashid, J., Ahmad, I., Almotiri, S.H., AlGhamdi, M.A., Masood, K. (2021). Racial identity-aware facial expression recognition using deep convolutional neural networks. Applied Sciences, 12(1): 88. https://doi.org/10.3390/app12010088

[6] LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature, 521(7553): 436-444. https://doi.org/10.1038/nature14539

[7] Arowolo, M.O., Ogundokun, R.O., Misra, S., Agboola, B.D., Gupta, B. (2023). Machine learning-based IoT system for COVID-19 epidemics. Computing, 105(4): 831-847. https://doi.org/10.1007/s00607-022-01057-6

[8] Ayo, F.E., Ogundokun, R.O., Awotunde, J.B., Adebiyi, M.O., Adeniyi, A.E. (2020). Severe acne skin disease: A fuzzy-based method for diagnosis. In International Conference on Computational Science and Its Applications, Cham: Springer International Publishing, pp. 320-334. https://doi.org/10.1007/978-3-030-58817-5_25

[9] Agarwal, S., Mahto, A.K. (2025). Skin cancer classification: Hybrid CNN-transformer models with KAN-Based fusion. arXiv preprint arXiv:2508.12484. https://doi.org/10.48550/arXiv.2508.12484

[10] Gupta, P., Vadgaonkar, N., Nirmal, J., Mehendale, N. (2025). A hybrid CNN-ViT framework for skin disease classification via feature extraction and selection. Neural Computing and Applications, 37: 27151-27177. https://doi.org/10.1007/s00521-025-11664-x

[11] Liu, R., Chen, Z., Zhang, P. (2025). Skin lesion classification based on ResNet-50 enhanced with adaptive spatial feature fusion. arXiv preprint arXiv:2510.03876. https://doi.org/10.48550/arXiv.2510.03876

[12] Qamar, S. (2025). Confidence-weighted semi-supervised learning for skin lesion segmentation using hybrid CNN-Transformer networks. arXiv preprint arXiv:2510.15354. https://doi.org/10.48550/arXiv.2510.15354

[13] Krishna, G.S., Supriya, K., Sorgile, M. (2023). Lesionaid: Vision transformers-based skin lesion generation and classification. arXiv preprint arXiv:2302.01104. https://doi.org/10.48550/arXiv.2302.01104

[14] Maheshselvi, T., Bharathiraja, V., Bragadheesh, R., Harishvijayabaskaran, S. (2025). AI-powered skin cancer detection using a CNN-transformer hybrid model. American Journal of Psychiatric Rehabilitation, 28(5): 204-212. https://doi.org/10.69980/ajpr.v28i5.357

[15] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639): 115-118. https://doi.org/10.1038/nature21056

[16] Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., Parmar, M. (2024). A review of convolutional neural networks in computer vision. Artificial Intelligence Review, 57(4): 99. https://doi.org/10.1007/s10462-024-10721-6

[17] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90. https://doi.org/10.1145/3065386

[18] Anggadwisunarto, A. (2018). Skin Cancer HAM10000 Hair Removal. https://www.kaggle.com/datasets/anggadwisunarto/skin-cancer-mnist-ham10000-hair-removal.

[19] Ogundokun, R.O., Li, A., Babatunde, R.S., Umezuruike, C., Sadiku, P.O., Abdulahi, A.T., Babatunde, A.N. (2023). Enhancing skin cancer detection and classification in dermoscopic images through concatenated MobileNetV2 and Xception models. Bioengineering, 10(8): 979. https://doi.org/10.3390/bioengineering10080979

[20] Kim, C., Jang, M., Han, Y., Hong, Y., Lee, W. (2023). Skin lesion classification using hybrid convolutional neural network with edge, color, and texture information. Applied Sciences, 13(9): 5497. https://doi.org/10.3390/app13095497

[21] Arshed, M.A., Mumtaz, S., Ibrahim, M., Ahmed, S., Tahir, M., Shafi, M. (2023). Multi-class skin cancer classification using vision transformer networks and convolutional neural network-based pre-trained models. Information, 14(7): 415. https://doi.org/10.3390/info14070415

[22] Bay, H., Ess, A., Tuytelaars, T., Van Gool, L. (2008). Speeded-up robust features (SURF). Computer Vision and Image Understanding, 110(3): 346-359. https://doi.org/10.1016/j.cviu.2007.09.014

[23] Tahir, M., Naeem, A., Malik, H., Tanveer, J., Naqvi, R.A., Lee, S.W. (2023). DSCC_Net: Multi-classification deep learning models for diagnosing skin cancer using dermoscopic images. Cancers, 15(7): 2179. https://doi.org/10.3390/cancers15072179

[24] Ibrahim, A.T., Abdullahi, M., Kana, A.F.D., Mohammed, M.T., Hassan, I.H. (2025). Categorical classification of skin cancer using a weighted ensemble

of transfer learning with test time augmentation. Data Science and Management, 8(2): 174-184. https://doi.org/10.1016/j.dsm.2024.10.002

[25] Qasim Gilani, S., Syed, T., Umair, M., Marques, O. (2023). Skin cancer classification using deep spiking neural network. Journal of Digital Imaging, 36(3): 1137-1147. https://doi.org/10.1007/s10278-023-00776-2

[26] Zhou, S., Tian, S., Yu, L., Wu, W., Zhang, D., Peng, Z., Wang, J. (2023). FixMatch-LS: Semi-supervised skin lesion classification with label smoothing. Biomedical Signal Processing and Control, 84: 104709. https://doi.org/10.1016/j.bspc.2023.104709

[27] Zafar, M., Amin, J., Sharif, M., Anjum, M.A., Mallah, G.A., Kadry, S. (2023). DeepLabv3+-based segmentation and best feature selection using slime mould algorithm for multi-class skin lesion classification. Mathematics, 11(2): 364. https://doi.org/10.3390/math11020364