

Article

HybridSkinLes: An Explainable CNN-and Transformer-Based Multi-Class Skin Lesion Classification Framework

May Issa Aldossary and Hina Gull *

Department of Computer Information Systems, College of Computer Science and Information Technology,
Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia;
mialdossary@iau.edu.sa

* Correspondence: hgull@iau.edu.sa

Abstract

Skin cancer is considered a deadly disease globally, and the timely identification of the disease may save human life. This research presents a CNN–Transformer-based fusion framework for automated multi-class skin lesion classification. This approach combines ResNet50 and Vision Transformer (ViT) to categorize skin lesions using the HAM10000 dataset. To assess their efficacy, a comparison with CNN and ViT models is also carried out. Seven classes of skin cancer are used for training the models, and class weighting is used to correct dataset asymmetry. According to the experimental dataset, the suggested hybrid framework shows improved performance over CNN and ViT, considering the accuracy (0.97) and macro-averaged F1-score (0.95). Furthermore, the efficiency of the suggested model is demonstrated by the fact that it delivers performance that is competitive with several existing approaches. Overall results indicate that hybrid CNN–Transformer architectures present a viable path for automated skin lesion categorization. Grad-CAM++ is integrated to enhance model understanding and promote medical confidence by enabling physicians to view visualizations that show the areas impacting the model’s conclusions. But there are still issues, including poor generalization, computational complexity, and a lack of external validation. Future research will concentrate on enhancing interpretability for practical implementation, integrating clinical information, and evaluating several datasets.

Keywords: skin cancer; deep learning; ResNet50; vision transformer; CNN; medical images

Academic Editor: Giorgio Maria
Di Nunzio

Received: 1 June 2026
Revised: 7 July 2026
Accepted: 13 July 2026
Published: 16 July 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The function of the skin is to safeguard us from different hazardous elements and physical traumas. A pigment called melanin has the capacity to filter out Ultraviolet radiation, shielding the skin from the harmful hazards of UV rays. The skin is divided into three layers, which are known as the epidermis, dermis, and hypodermis. The epidermis, the outermost covering, protects it from outside factors. The dermis is located beneath it and comprises connective cells as well as glands for sweating. The hypodermis is a fat-rich sublayer of the skin [1]. Skin problems are a major medical issue. They impact more individuals each year and are the root cause of medical appointments globally [2,3]. There have been reports of over three thousand skin conditions, both critical and chronic, affecting people from all social classes and generations [4]. Some of these can be deadly,

like skin cancer. Even in the event that this is untrue, they may nevertheless place a heavy financial and quality-of-life strain on people [5]. Skin lesions encompass multiple recurrent diseases globally, with over 3.3 million cases [6]. Melanoma is the deadliest form of skin cancer that has been found so far. The reason for this is that doctors and surgeons have a tough time detecting it in its early phases. In 2020, more than 300 thousand patients globally were confirmed to have this cancer [7]. The disease has proven fatal, although it is easily treatable if identified at an early stage. According to a study, when detected and treated early, the 5-year mortality-free probability of this fatal disease is 99%, while for late-stage melanoma, it is only 25%; thus, whether the illness can be identified early has a significant effect on patients [8]. The standard inspection procedure involves a physician's ocular evaluation followed by a patient's skin images to help in determining the condition [9]. However, the symptoms and traits of various types of skin cancers are highly similar, and the ambiguity in decision-making by doctors has resulted in many patients not being properly identified and treated in time [10]. The application of AI in diagnostics to offer a second opinion and aid medical experts in diagnosing skin cancer is becoming popular [11–18]. One helpful use of AI in the medical area is Computer-Aided Diagnosis and Detection (CAD), especially for screening skin lesions using dermoscopic pictures. Pre-screening for skin cancer is made easier by CAD systems, which avoids diagnostic delays. After diagnosis, the cases are sent for further examination to professional skin specialists. For skin cancer identification and classification, different models of AI, for instance, CNNs, are employed. CNNs categorize new, unseen instances using the relationships they have learnt from training using dermoscopic pictures and their ground truth [19–21]. Similarly, the Vision Transformer has also been used in various studies for training using dermoscopic images. Both of them have also been applied as standalone models for skin cancer classification. When it comes to collecting local characteristics like edges, textures, and color changes, CNNs (like ResNet50) thrive. They have trouble comprehending long-range relationships in a picture, though. Vision Transformers, on the other hand, are less successful at learning fine-grained features but are able to gather broad contextual information. Understanding both local and global features is necessary for clinical visual interpretation, especially for the categorization of skin cancer. Consequently, to get around the drawbacks of standalone CNN and Transformer structures, hybrid models are required [22].

The layout of this paper is organized into the following sections. Section 2 discusses the previous studies. In Section 3, the hybrid model is presented. Methodology is given in Section 4. The findings are discussed in Section 5. Section 6 illustrates the explainability of the model. Finally, the limitations of the study and the conclusions are indicated in Sections 7 and 8, respectively.

2. Related Work

Skin cancer detection provides an appealing subject area for the use of artificial intelligence in image-based diagnostics. AI-based algorithms can classify clinical or Epiluminescence microscopy images with amazing accuracy after analyzing enormous datasets [23]. Several studies showed the implications of AI algorithms in the successful detection of skin cancer. Moreover, timely identification of skin cancer is critical since it allows for the most effective treatment. However, because of the resemblance of different kinds of tumors in terms of appearance, shape, hue, and dimensions, identifying skin cancer is hard to identify. Using training photos, Ahmad et al. [24] used a deep learning strategy for skin lesion classification by merging and modifying three initially trained deep learning-based architectures. Furthermore, the class imbalance is handled to improve the lesion categorization. The investigational results show that the proposed framework yielded encouraging outcomes. Additionally, Hasan et al. [25] introduced

DermoExpert, which was assessed on three datasets. By increasing classification accuracy, it beat conventional models by combining segmentation, augmentation, and transfer learning with CNN. Nevertheless, the augmentation procedure may produce artifacts that were absent from the initial clinical settings, resulting in a biased model. Jinnai et al. [26] collected medical photos of skin lesions with pigmentation from 3551 individuals participating in a study, which included malignant tumors. By choosing people at random and one image per patient, they created a test set. A proposed model was then trained and evaluated. Additionally, ten trainees and ten specialized dermatologists had their diagnosis accuracy compared to the automated model. The skin cancer was effectively detected by the system. Thomas et al. [27] applied interpretable deep learning algorithms to the most common skin malignancies. The authors reveal that the majority of derma work is agreeable to automated analysis because distinct cancer types account for more than 90% of diagnoses. This work focuses on characterizing the tissue by splitting it into 12 significant dermatological classifications, including the roots of hair and sweat glands, as well as distinguishing the skin's coatings. The network is trained to represent the problem domain in the same way that a pathologist would. Therefore, the results are very exact. Another approach suggested by Jojoa et al. [28] consists of two stages: a starting point that consequently crops the region that is important within a dermatoscopic picture using neural networks, and a second stage that classifies lesions using a ResNet152 structure. As a result, both the sensitivity and specificity receive a high score (higher than 0.8), indicating that the algorithm is good for correctly distinguishing between malignant and benign tumors and has no prejudice towards any of them. Similarly, a CNN model was tested by Selvi [29] to categorize skin lesions. The correctness of the model was 91.51%. The efficacy of the model was assessed against comparable investigations in the literature, and it was discovered that it performed better than most studies. The model was tested by seven physicians. In the first phase, it was determined that the model could accurately diagnose skin lesions in practice with 90.28% accuracy. The model rectified the experts' incorrect diagnoses by 11.14% in the subsequent stage. Different approaches using neural networks for identifying and categorizing skin cancer were examined in a study. All of these procedures are harmless. To detect skin cancer, it is necessary to first prepare and segment images, then extract features, and finally categorize them.

Similarly, the Transformer has also been used to classify skin cancer widely. Adablanu et al. [30] integrates spatial sequence modeling, CNN, and ViT. Gaussian blur, normalization, and augmentation were used in preprocessing to overcome class inequity and minimize noise. The public HAM10000 dataset was trained using a CPU and Google Colab environment. Under the same preprocessing conditions, the performance of VINCE-NET was compared to CNN, ResNet-50, and convolutional networks (DenseNet-121/201). Xin et al. [31] suggested a skin cancer categorization network based on ViT and found that the network performs satisfactorily in skin cancer classification. Additionally, their research used several of ViT's features to further mine ViT and enhance classification performance. The study does multi-scale patch embedding and serializes the picture to maximize the differences in the coding outputs of different data. It proved to be 94.3% reliable on the publicly available dataset and 94.1% on their medical dataset. By utilizing both Transformers and CNNs, the Swin Transformer model is suggested by Ayas [32]. Additionally, important elements of the suggested method are thoroughly examined to guarantee an effective process for the categorization of skin lesions. The suggested approach achieves sensitivity, specificity, accuracy, and balanced accuracy values of 82.3%, 97.9%, 97.2%, and 82.3%, respectively, after being thoroughly tested on the ISIC dataset. According to experimental data, the suggested approach surpasses the majority of existing cutting-edge techniques in multi-class skin lesion categorization and has the most balanced accuracy value. Cai et al. [33] developed CDMT-Unet, a revolutionary

hybrid model that uses the Mnet as well as the Transformer for accurate skin lesion segmentation of pictures and detection. This model combines an attention mechanism that concentrates on more relevant information with the advantages of the Haar wavelet transform, which can capture global texture and edge information through multi-scale analysis. Combining these approaches allows the model to fully define the skin lesion region by utilizing both low-frequency structural data and high-frequency detail. The proposed framework is the combination of ViT, CNN, and Unet architecture. According to experimental results, the proposed model performs exceptionally well on two publicly available skin cancer datasets.

Moreover, hybrid techniques combining CNN and Transformer have also been used to detect skin cancer. To accurately classify dermoscopic skin lesion pictures, the work [34] presents a unique architecture called Enhanced Vision Transformer (EViT) with Dens169. The suggested architecture combines DenseNet169 with EViT to make use of both fine-grained local information and global context. Six attention-based encoder blocks consist of a multihead self-attention (MHSA) mechanism and layer normalization, forming the EViT Encoder component, which facilitates effective global spatial comprehension. They added a Spatial Detail Enhancement Block (SDEB) with three parallel convolutional layers and a fusion layer to maintain the local spatial continuity lost during patch segmentation. The edge, boundary, and texture details—all essential for lesion detection—are rebuilt by these layers. Modified for dermoscopic data, the DenseNet169 backbone captures local characteristics that enhance global attention features. With an accuracy of 97.1% and a sensitivity of 90.8%, the suggested hybrid approach is better than the traditional techniques on the ISIC 2018 images. By effectively combining the global and local aspects, the hybrid framework offers a reliable method for the timely identification of skin lesions. ThangaPumi [35] created a hybrid model based on CNN, ARP and ViT. Next, they created a more sophisticated classification model that can classify various categories of skin cancer accurately using this feature representation. To solve the issue of underrepresented classes and picture variability, the proposed approach incorporates a thorough preprocessing pipeline, which includes balancing and improving the dataset. They tested their algorithm using the publicly accessible HAM10000 dataset. With an overall reliability of 98.2%, precision of 0.94, recall of 0.96, and macro-F1-score of 0.95, the ARP-ViT-CNN model outperformed the traditional CNN-only and Transformer-only models on the dataset, as expected. Agarwal et al. [36] investigated Convolutional Kolmogorov–Arnold Network (CKAN)-based Sequential and Parallel Hybrid CNN–Transformer models. With CNNs extracting local spatial features, transformers modeling global dependencies, and CKAN facilitating nonlinear feature fusion for enhanced learning about representation, their method combines transfer learning with significant data augmentation. They test their models on several available datasets with different data distributions and class imbalances in order to assess their applicability. According to experimental findings, hybrid CNN–Transformer designs increase classification efficacy by efficiently capturing both spatial and contextual variables. Furthermore, the use of CKAN produces less biased representations by improving feature fusion through learnable activation methods.

3. Proposed Model

This study suggests a deep learning-based model called HybridSkinLes for automated skin lesion classification using a hybrid CNN ResNet50 and Vision Transformer technique. The suggested model emphasizes enhancing diagnostic performance through strong training techniques and architectural variety. These separately pre-trained backbone models (CNN and ViT) are fused at the feature level. ResNet50 was chosen for its ability to accurately capture local picture properties such as

lesion borders, color changes, texture patterns, and fine-grained dermatological structures [37]. This allows the model to learn both global contextual linkages and spatially localized patterns at the same time. DF, VASC, AKIEC, BKL, BCC, NV and MEL are the seven skin lesion classifications that are predicted by the proposed model. The complete architectural details of the proposed model are explained in the following section.

4. Methodology

4.1. Dataset and Preprocessing

The publicly available HAM10000 image dataset [38] from Kaggle, which includes multi-class skin lesion pictures, is used to assess the suggested experimental framework. One of seven diagnostic categories is linked to each picture. With 80% of the data utilized for training and 20% for validation, the dataset is divided into training and validation sets using a stratified sampling technique to maintain class distribution. To guarantee compatibility with deep learning systems, every picture is scaled to 224×224 pixels. Standard deviation values and ImageNet mean are used for standard normalization. Data augmentation techniques, such as random rotation ($\pm 20^\circ$) and random horizontal flipping, are added to the training set to increase generalization and decrease overfitting.

4.2. Proposed Model Architectures

Three deep learning models were implemented and assessed under similar experimental settings (i.e., the same data splits, training hyperparameters, input resolution of 224×224 pixels, and transfer-learning procedure) in order to provide a thorough and equitable comparison. The timm (PyTorch Image Models v1.0.28) package was used to instantiate each model, and ImageNet-1k pre-trained weights were used for initialization. Each model's final classification head was swapped out for a new fully connected layer that produces seven classes, i.e., akiec, bcc, bkl, df, mel, nv, and vasc, that match the skin lesion classifications in the HAM10000 dataset. This transfer-learning method adapts the models to the domain-specific goal of dermoscopic image classification while utilizing rich visual data acquired from millions of real pictures. Convolutional and Transformer encoding algorithms are combined in a number of recent studies to analyze medical images. Manzari et al. [39] utilize an effective convolution process to build the proposed attention mechanism. Additionally, they try to learn smoother decision boundaries in order to mitigate their Transformer model's vulnerability to adversarial attacks. In order to achieve this, they permute the feature mean and variance within small matrices to enhance the shape characteristics of a picture in the high-level feature domain. Similarly, Nanni et al. [40] present new variations of Adam. They used these novel techniques in conjunction with multiple CNN topologies to create an ensemble of classifiers that works better than stochastic gradient descent and other Adam-based techniques. Furthermore, the work integrates an ensemble of Transformers based on several topologies, including Deit, Vit, Swin, and Coat, with the ensemble of CNNs. Moreover, Shao et al. [41] developed a Transformer-based algorithm that investigated spatial and morphological data. With excellent comprehension, their suggested model can handle binary/multiple and unbalanced/balanced classification. Our model differs in the following ways: instead of using cross-attention or component-wise addition methods, we transform both branches into a common 512-dimensional space prior to combining them. Moreover, the model also addresses the imbalance weight of the HAM10000 dataset with end-to-end fine tuning. Additionally, since the design does not require any unique attention modules, it may be easily reimplemented using any contemporary deep learning package. Complete baseline details of all three architectures are given below.

4.2.1. Baseline Architecture CNN ResNet50

ResNet50, a popular CNN-based model for medical image classification, serves as the foundation for the model design [42]. ResNet50 was chosen because of its proven performance in medical image applications, especially in dermatology, where it has been established that transfer learning from ImageNet greatly improves feature extraction in dermoscopic pictures. This selection is appropriate for the datasets examined here because it balances model depth and computational efficiency and has demonstrated high performance in previous melanoma-detection investigations [43]. The architecture begins with a 7×7 convolutional layer. Max pooling comes after a 7×7 convolutional layer in the design. Four hierarchical residual stages with bottleneck blocks make up the main structure. A series of 1×1 , 3×3 , and 1×1 convolutions is included in each bottleneck block to enable effective feature extraction while preserving computational efficiency. One way to express the network is

$$y = F(x) + x \quad (1)$$

where the residual mapping acquired by stacked convolutional layers is represented by $F(x)$. Global average pooling is applied to the ultimate depiction, and then a fully linked layer is used for categorization [44]. In the proposed framework, the initial classification layer of ResNet50 is removed, and the network produces a 2048-dimensional feature representation for each input image.

If we consider an input image as I , ResNet50 will extract the features as follows:

$$\text{Feature (CNN)} = \text{ResNet50} (I) \quad (2)$$

where Feature (CNN) belongs to V^{2048} .

4.2.2. Baseline Architecture Vision Transformer (ViT)

Computer vision has undergone a revolution thanks to a neural network design called a Vision Transformer (ViT) [45]. Input photos are processed by ViT sequentially by splitting into same-sized chunks of patches. These patches are then linearly embedded, followed by using self-attention processes to capture global reliance, in contrast to conventional convolutional neural networks (CNNs) that process images pixel by pixel. Token layouts are created by linearly embedding fixed-size patches of 16×16 pixels from each input picture. To maintain spatial information, a learnable positional encoding is included. The Transformer encoder is made up of several identical layers, each of which has

- i. Multi-Head Self-Attention (MHSA);
- ii. Feed-Forward Network (MLP);
- iii. Residual connections and layer normalization.

In the proposed architecture, the Vision Transformer is used to record global contextual data and long-range spatial connections. The input picture is split up into non-overlapping patches, which are then processed via many self-attention layers after being converted into embedding vectors. The ViT creates 768-dimensional features.

If we consider an input image as I , ViT will extract the features as follows:

$$\text{Feature (ViT)} = \text{ViT} (I) \quad (3)$$

where Feature (ViT) belongs to V^{2048} .

4.2.3. Proposed Hybrid Model

As mentioned earlier, the proposed hybrid model HybridSkinLes combines Transformer-based attention mechanisms with convolutional feature extraction. After feature extraction, both feature representations are mapped onto a shared 512-

dimensional latent space using two separate projection layers, as the feature dimensions of both CNN and ViT differ substantially. The following equations show the computation mechanism of projected CNN and ViT features by applying the Rectified Linear Unit activation function:

$$\text{Project(cnn)} = \text{Dropout}_{0.3}(\text{ReLU}(\text{W}_{\text{cnn}}\text{Feature}(\text{CNN}) + \text{bcnn})) \quad (4)$$

$$\text{Project(ViT)} = \text{Dropout}_{0.3}(\text{ReLU}(\text{W}_{\text{vit}}\text{Feature}(\text{ViT}) + \text{bvit})) \quad (5)$$

where W_{cnn} and W_{vit} are the weight learning matrices, and a dropout rate of 0.3 is applied to reduce overfitting. The following equation shows the result:

$$\text{Project(cnn)}\text{Project(ViT)} \text{ belongs to } V^{512} \quad (6)$$

The concatenation operation is used to create the fusion as follows:

$$\text{Fusion} = \text{Concatenate}(\text{Project(cnn)}, \text{Project(ViT)}) \text{ belongs to } V^{512} \quad (7)$$

Then, a fully linked classification module receives the fused feature vector with ReLU activation and a dropout rate of 0.3. Finally, the softmax function is applied for the prediction of seven classes:

$$\text{Hfusion} = \text{Dropout}_{0.3}(\text{ReLU}(\text{W}_{\text{fusion}}\text{Fusion} + \text{bfusion})) \quad (8)$$

$$\text{Predicted Class} = \text{Softmax}(\text{W}_{\text{predicted}}\text{Hfusion} + \text{bpredicted}) \quad (9)$$

A detailed representation of HybridSkinLes is given in Figure 1.

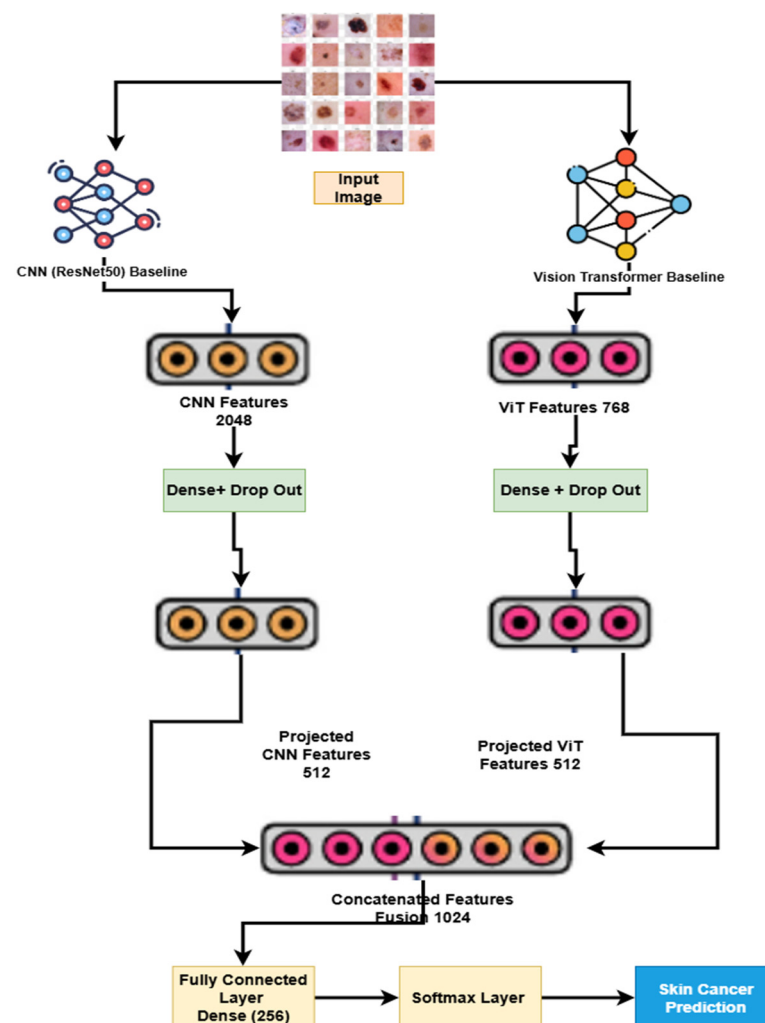


Figure 1. Detailed architecture of HybridSkinLes.

4.3. Model Training

A cross-entropy loss function, considered typical for multi-class classification tasks, was used to train the model. Optimization was done using Adam's optimizer. To balance convergence speed and stability during training, a learning rate of 1×10^{-4} is used. The models are trained across ten epochs with a batch size of thirty-two. As both architectures, ResNet50 and ViT, have pre-trained characteristics and are derived from ImageNet weights [46], 10 epochs were selected. Small batch sizes, such as 32, are appropriate for computers with less power because they need less memory. For data such as HAM10000, which has a severe class imbalance, batch sizes less than 16 result in large variability in slope estimates; batch sizes greater than 64 increase memory requirements without producing appreciable precision gains on this dataset size. Moreover, a learning rate of 1×10^{-4} adheres to standard procedure for fine tuning on medical images [47]. Additionally, each model is trained separately using the same stratified train-validation split, guaranteeing that variances in performance are only due to architectural changes rather than inconsistent data.

4.4. Evaluation Protocol

The suggested models were evaluated on the testing dataset following the training procedure. The performance of the architecture was assessed using the metrics given below in Figure 2:

$$\text{Accuracy} = \frac{\text{No. of correctly classified samples}}{\text{No. of all samples}}$$

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

$$F - 1 \text{ Score} = 2X \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$F - 1 \text{ Score} = \frac{2X TP}{2X TP + FP + FN}$$

Figure 2. Model evaluation metric.

The receiver operating characteristic (ROC) curve graphically depicts the relationship between true and false positives of a classifier. The area under the curve (AUC) of the probability curve illustrates the ability of the model to distinguish across classes. Better class separation by the classifier is indicated by a higher AUC value. The model's performance is also analyzed statistically by using the McNamar test.

5. Results and Discussion

This section will highlight the findings of our study.

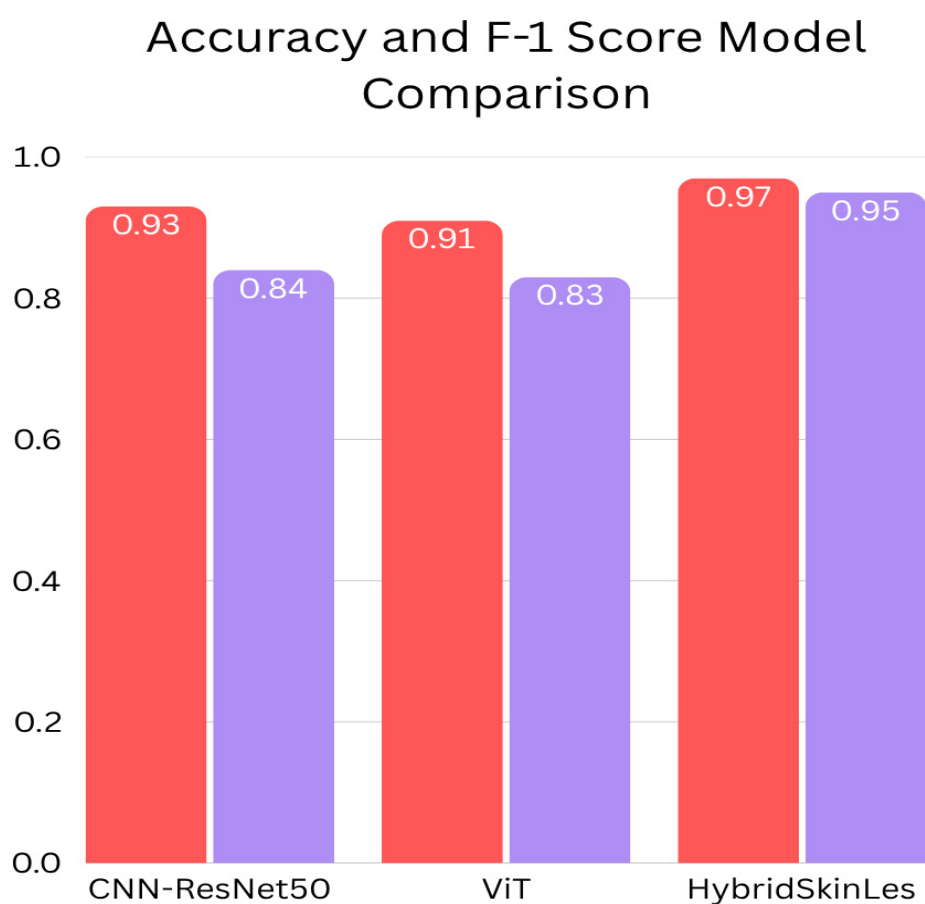
5.1. Accuracy and F1-Score

The performance of all three models in terms of their accuracy and F1-score is given in Table 1.

Table 1. Accuracy and F1-score.

Model	Accuracy	F1-Score
CNN-ResNet50	0.93	0.84
ViT	0.91	0.83
HybridSkinLes	0.97	0.95

With overall classification efficacy, the hybrid CNN–Transformer model fared better than the Vision Transformer and CNN models. Although the CNN model performed well in collecting local texture information, it had trouble differentiating visually comparable classes. Conversely, the Vision Transformer used global contextual information, but because of data constraints, it performed rather poorly. The hybrid model produced better classification results for all classes by successfully combining both local and global feature representations. Its improved management of the class imbalance found in the HAM10000 dataset is demonstrated by its greater accuracy and larger F1-score. The comparison of all three models is given in Figure 3 below:

**Figure 3.** Accuracy and F1-score comparison. Red represents accuracy and purple represents F-1 score.

The HAM10000 dataset has seven distinct skin lesion classifications, which are used in Table 2 to compare our suggested hybrid model with cutting-edge techniques. For classification tasks, accuracy is regarded as a crucial performance parameter, and our model performed equal to or better than other cutting-edge techniques.

Table 2. Comparison of HybridSkinLes with cutting-edge techniques.

Reference	Accuracy Score
Pacal et al. [48]	0.93
Yang et al. [49]	0.94
Xin et al. [31]	0.94
George et al. [50]	0.95
Najjar et al. [51]	0.97
Aboulmira et al. [52]	0.95
Al Shafi et al. [53]	0.96
HybridSkinLes	0.97

5.2. Confusion Matrix

The CNN, Vision Transformer (ViT), and hybrid models' classification performance across the seven skin lesion classes is shown in the confusion matrices. The distribution of correctly identified samples along the diagonal shows a substantial difference. With the most focused diagonal, the hybrid model exhibits better classification accuracy for both majority and minority classes. The CNN and ViT models, on the other hand, show more distributed off-diagonal values, which indicate greater misclassification rates, especially across visually identical lesion types.

With a near-perfect classification rate (0.99), the CNN model (Figure 4) performs well in dominating classes like melanocytic nevus (nv) and vascular lesion (vasc), demonstrating its efficacy in learning local texture information. However, it performs worse in minority and clinically critical classes. In a similar vein, the model only achieves about 0.43 correct classification for dermatofibroma (df), suggesting that it has trouble collecting discriminative characteristics for uncommon classes. These findings imply that although CNNs work well for dominating patterns, they are not robust enough to handle modest inter-class differences and class imbalance. Similarly, the Vision Transformer (Figure 5) enhances performance for specific courses. For example, compared to the CNN model, melanoma classification increases to about 84%. Furthermore, there is an improvement in the Actinic Keratosis (AKIEC) categorization. Nevertheless, additional kinds of faults are introduced by the ViT model. Basal Cell Carcinoma (bcc), for instance, has a lower classification rate (81%) and is clearly confused with other groups. Additionally, dermatofibroma (df) is still poorly categorized (48%), suggesting that minority classes continue to be a challenge for the model. Additionally, the ViT model shows confusion between bkl, mel, and nv, indicating that although global attention is helpful, it may miss fine-grained local cues that are essential for differentiating related lesions.

By successfully fusing local feature extraction with global attention processes, the hybrid model performs noticeably better than both CNN and ViT (Figure 6). The significant diagonal dominance in every class makes this clear. With a significant decrease in misclassification, the categorization of melanoma (mel) rises to around 91%. Performance for dermatofibroma (df) significantly increases to 91%. Notable improvements are also shown in BKL and AKIEC, which reached around 95% and 82%, respectively. For vascular lesions (vasc) (100%) and nevus (nv) (98%), the model continues to perform quite well. Furthermore, there is much less inter-class confusion, especially between nevus and melanoma, which is essential for clinical dependability.

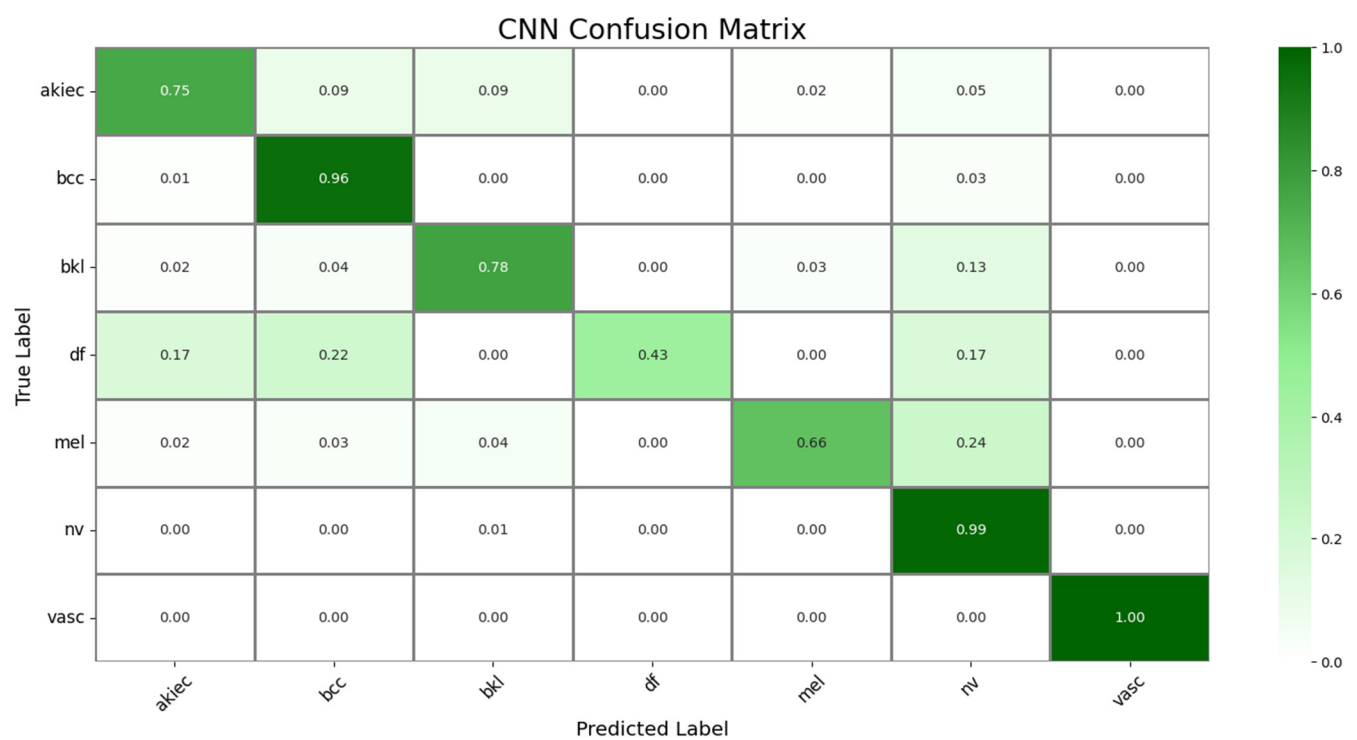


Figure 4. CNN confusion matrix.

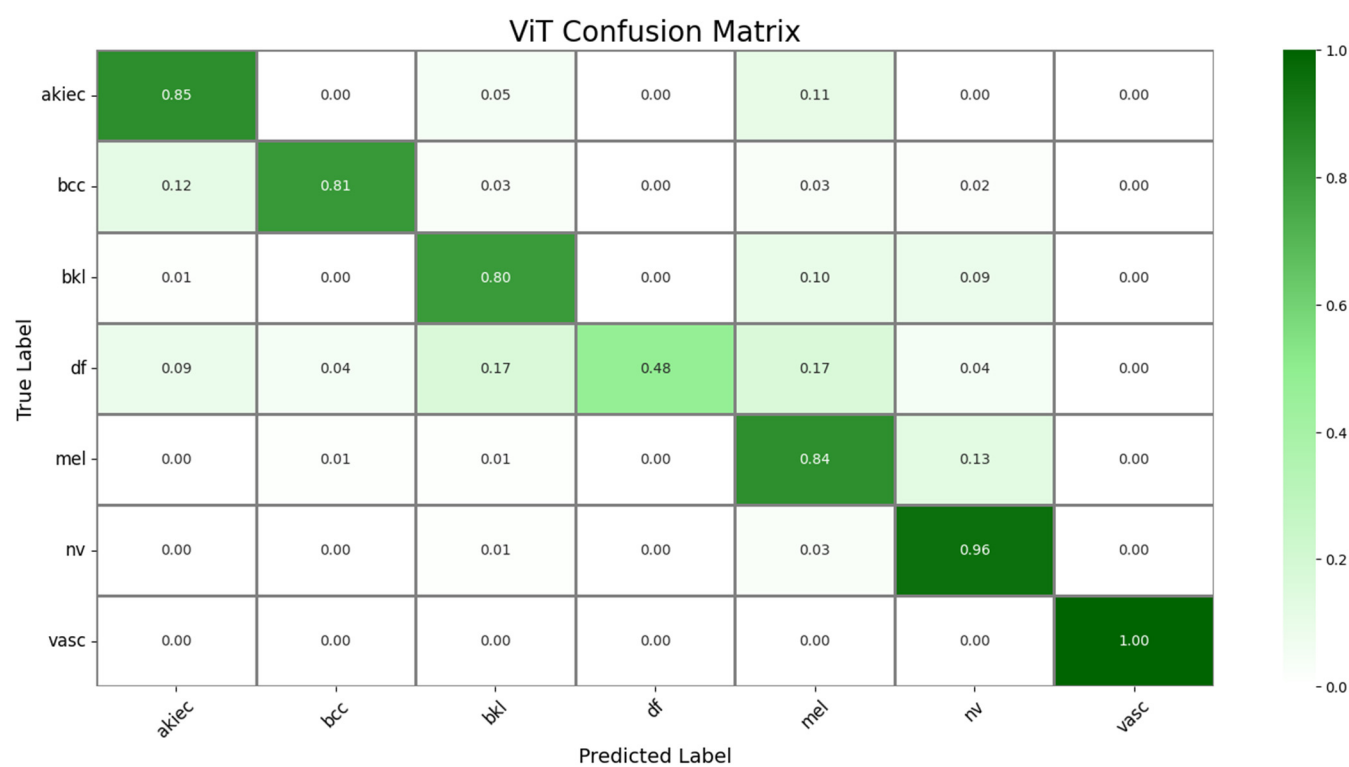


Figure 5. ViT confusion matrix.

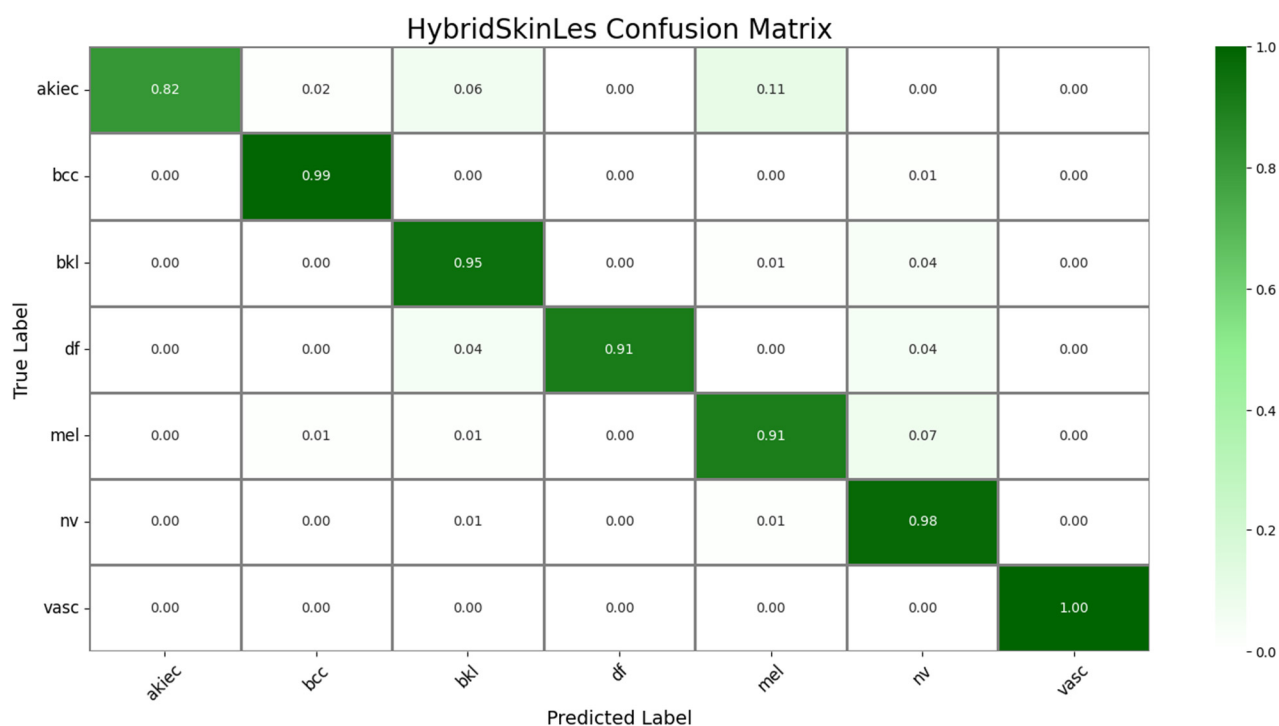


Figure 6. HybridSkinLes confusion matrix.

5.3. ROC Curve Analysis

The effectiveness of classification across all decision levels by all three models is shown by the ROC curve analysis. All three models generate arcs that are clearly above the diagonal requirement, which indicates substantial differentiated capability. With an AUC of 0.995, the CNN model performs exceptionally well in class differentiation. Transformer-based architectures are similarly quite successful for this job, as seen by the ViT model's equivalent AUC of 0.994. With the greatest AUC of 0.998, the hybrid model, however, performs better than both methods, demonstrating higher overall classification capabilities. One important finding is that, especially in areas with low false positive rates, the ROC curve of the hybrid model regularly stands above that of CNN and ViT (Figure 7). This implies that the hybrid model minimizes false alarms while maintaining greater true positive rates, which is crucial in medical diagnosis settings where lowering false positives and false negatives is crucial.

5.4. Per-Class Accuracy Analysis

This section shows the performance of HybridSkinLes across distinct skin lesion classes in the per-class accuracy analysis (Figure 8). With nearly 100% accuracy, the CNN model performs well for dominating classes like nevus (nv) and vascular lesions (vasc). However, it performs badly on minority or difficult classes like melanoma (mel) (66%) and dermatofibroma (df) (44%), demonstrating problems in addressing modest feature variations and class imbalance. For challenging classes, the ViT model outperforms CNN. Specifically, accuracy for melanoma improves to about 84% and for df rises to around 48%, demonstrating the benefit of global feature modeling through self-attention processes. But when compared to CNN, ViT performs somewhat worse in several classes, such as bcc and bkl. In every class, the hybrid model continuously attains better precision. It greatly enhances effectiveness in difficult groups, with melanoma accuracy over 90% and df reaching almost 91%. Meanwhile, it continues to work almost flawlessly for dominating classes like vasc and nv. This balanced improvement indicates that the hybrid approach successfully addresses both inter-class similarity and class imbalance.

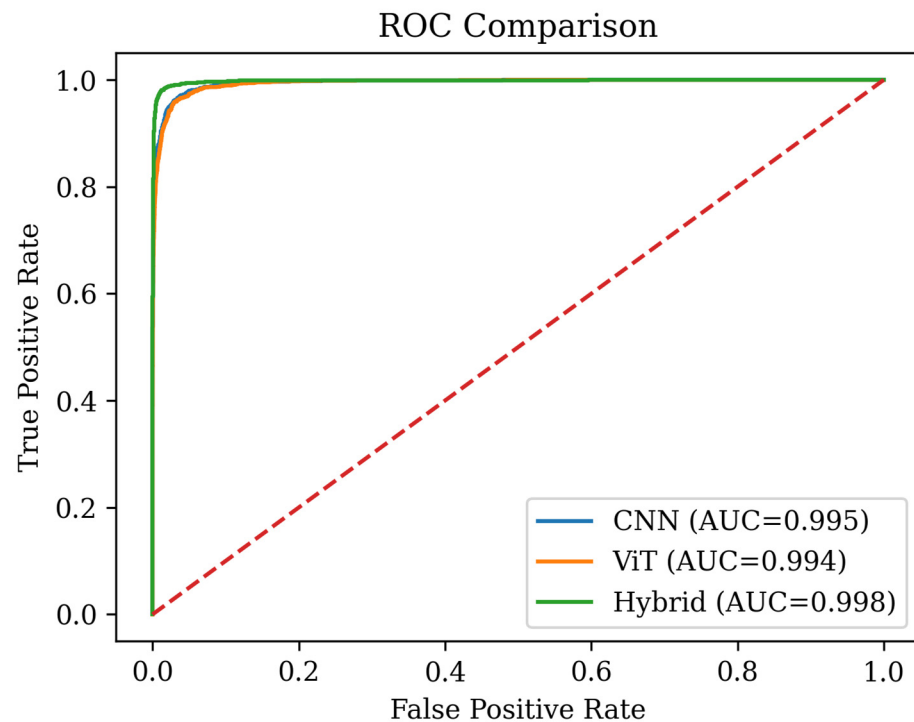


Figure 7. ROC comparison.

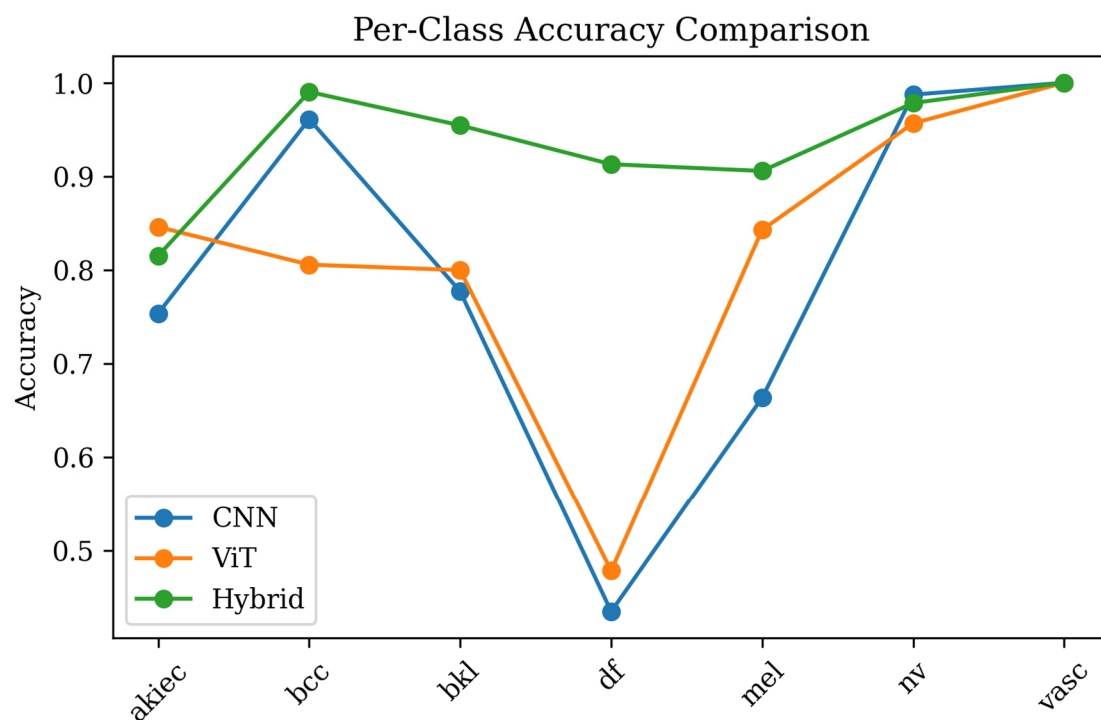


Figure 8. Per-class accuracy comparison.

Moreover, the following figures (Figures 9–11) show the per-class ROC curve for all three models. Figures indicate all FPRs (false positive rates) and TPRs (true positive rates) along with per-class accuracy. The CNN model achieved the highest performance (1.00) in classes such as bcc and df, while the lowest performance was shown in bkl and mel. The ViT model performed highest (1.00) in classes such as bcc, df, and vasc, while showing

average performance in other classes, whereas HybridSkinLes achieved the highest performance (1.00) in classes such as akiec, bcc, df, nv and vasc., HybridSkinLes classifies the other two classes, mel, and bkl, much better (0.99) than the other two architectures. Notably, although there is class imbalance in the original data, melanoma (mel), the most important clinical class because of its aggressiveness and complexity in diagnosis [54], was categorized with an AUC of 0.99, suggesting a substantial distinction between cancer and non-cancerous instances.

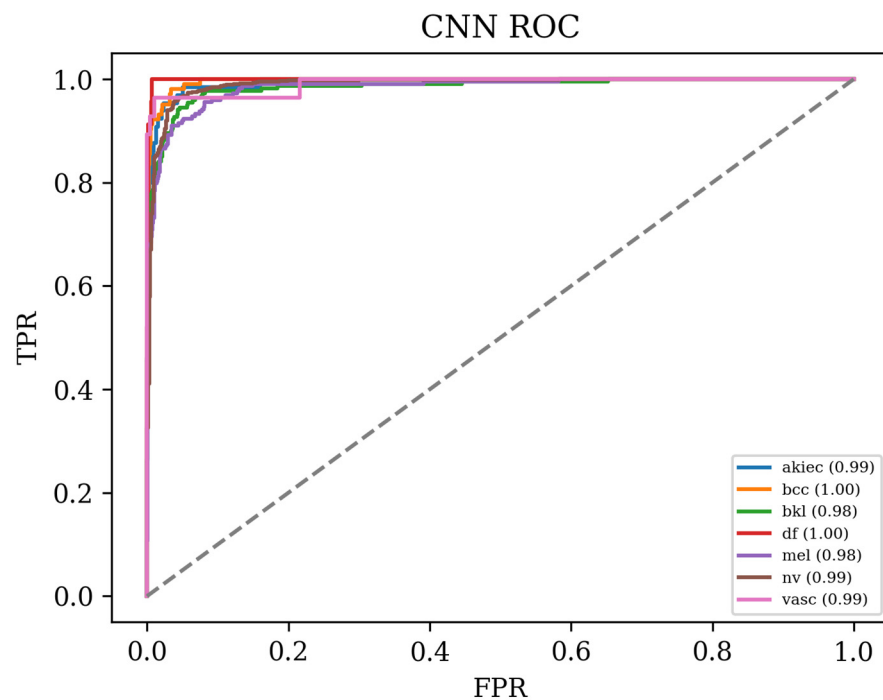


Figure 9. CNN ROC curve.

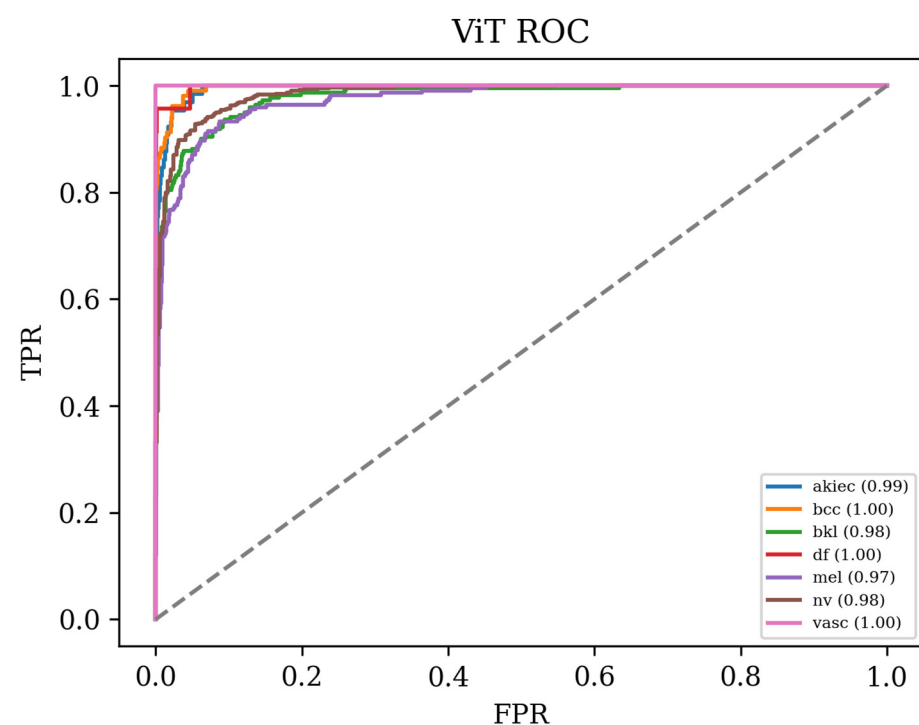


Figure 10. ViT ROC curve.

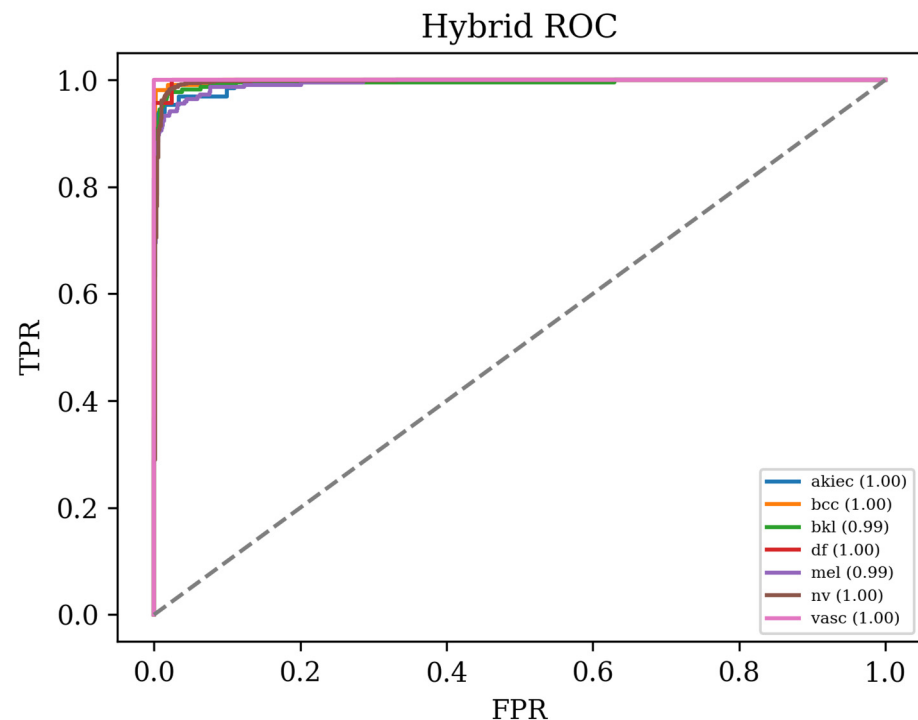


Figure 11. HybridSkinLes ROC curve.

5.5. Statistical Significance

The statistically significant nature of performance variations across the models was assessed using McNemar's test. When comparing two categorization models, McNemar's test considers the samples where the models' predictions diverge. It assesses the statistical significance of the outcome discrepancy between two models [55].

The discordant predictions (those in which one model is true and the other is incorrect) are considered when creating a contingency table for McNemar's test. It is based on image-level predictions. The following is the categorization of sample data:

n00: no. of samples wrongly predicted by both models.

n01: no. of samples wrongly predicted by model A but not model B.

n10: no. of samples wrongly predicted by model B but not model A.

n11: no. of samples correctly predicted by both models.

Null Hypothesis: $n_{01} = n_{10}$.

For our experiment, the test was performed on the same dataset that is used for validation. Several lesion groups had very few samples, so a class-wise test was not carried out. McNemar's test basically represents a paired chi-square test, so we use the following formula (Yates' continuity correction) to get the test value [56].

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$

Here three pairwise comparisons are conducted; the probability of finding false-positive results increases, so we apply Bonferroni Correction [57] to adjust the p value as follows:

$$P - \text{value (adjusted)} = p - \text{value} \times k \quad (10)$$

where k is the number of comparisons.

Test Results:

We set the null hypothesis as follows:

H0: The performance of the two models under comparison is identical or shows no difference.

Due to the same error patterns of both models, the comparison of CNN and ViT produced a p -value of 0.7999, suggesting no statistically significant difference. With p -values less than 0.001, the hybrid model, on the other hand, demonstrated statistically significant improvements over both CNN and ViT. In particular, there were significantly more samples properly identified by the hybrid model than incorrectly labeled by CNN and ViT. These findings demonstrate the hybrid architecture's enhanced classification capabilities by confirming that the performance increases are statistically significant and not the product of random fluctuation. The results are given in Figure 12 below:

Model 1	Model 2	n11	n10	n01	n00	χ^2	P value	P-value (adjusted)	Hypothesis Results
CNN	ViT	1717	130	110	46	1.504167	0.220031364	0.660094091	Accept
CNN	Hybrid	1791	56	131	25	29.28342	6.25283×10^{-8}	1.87585×10^{-7}	Reject
ViT	Hybrid	1783	44	139	37	48.28415	3.6872×10^{-12}	1.10616×10^{-11}	Reject

Figure 12. Hypothesis acceptance/rejection.

6. Deep Learning Explainability and Interpretability of Framework

Deep learning systems must offer some explainability in order to be adopted by physicians, who need to understand how a model makes judgments [58]. To improve this understanding of DL models, many explainable artificial intelligence (XAI) strategies have been developed in recent years. These approaches fall into four general categories: perturbation-based (such as SHAP, LIME, Occlusion Sensitivity), Gradient-Based (such as Integrated Gradient, Smooth Gradient, CAM, Grad-CAM, Grad-CAM++), Decomposition (LRP), and attention methods (Vision Transformer). The literature showed that for image categorization, methods like CAM and Grad-CAM work well and achieve high levels of precision. While IG is used for classification with remarkable ROC scores, LRP is known for its precision in tasks involving segmentation. While perturbation-based approaches evaluate model resilience, attention-based approaches enhance performance and interpretability by concentrating on pertinent regions. Trainable attention models dynamically improve feature focus, whereas LIME offers insights that are independent of models. We have selected the specific branches of CAM, Grad-CAM, and Grad-CAM++, as they have shown the best precision for image-based classifications [59]. Among these three techniques, the explainability capabilities of CAM are only present in CNNs with GAP at the penultimate layer, which is a fully connected invisible layer. Furthermore, CAM requires retraining several linear classifiers after training the fundamental model. The Grad-CAM approach presents the concept of reverse propagation to address this issue. It also considers a GAP of the partial derivatives to determine the weight independent of the placement of a specific activation map. Even though related characteristics are confined to a small pixel region, the Grad-CAM++ approach produces more emphasized heatmaps than Grad-CAM [60]. So, after comparison, by providing brief visual descriptions for the categorization of skin lesions, this study also enhances the readability of deep learning-based frameworks using Grad-CAM++. In this case, Grad-CAM++ produced consistent, visual mappings that correspond with model decisions and offered explanations. The results showing the explainability of the model utilizing Grad-Cam++ are shown in Figure 13. The left side of the image is the original image taken from the HAM10000 dataset, while the right side shows a heatmap visualizing the skin cancer region.

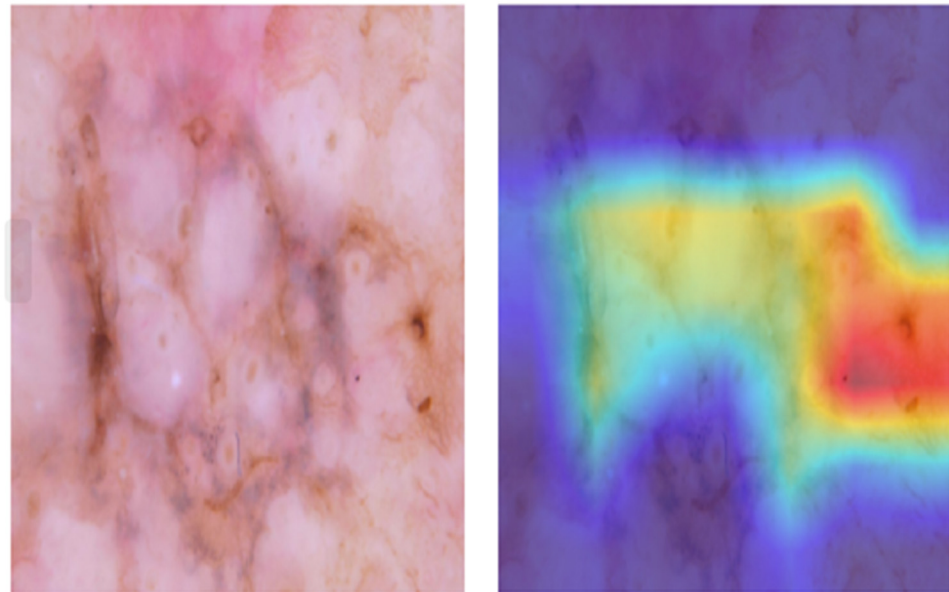


Figure 13. Grad-Cam++ visualization.

Moreover, Figures 14–16 display heatmaps for each of the seven classes. They display both the original image and Grad-CAM renderings for each of the seven classes. The heatmaps are constantly aligned with clinically significant lesion characteristics. CAM's interpretative capacity is still restricted, despite the fact that it provides a helpful model; the ability to concentrate on diagnostically significant locations is confirmed by the method of seeing the areas of an image that most affect a model's conclusion. The direct reliability of these approaches in clinical practice is questioned, since they might occasionally produce deceptive artifacts or emphasize areas that are not actually related to the underlying disease. Therefore, rather than being conclusive diagnostic evidence, these visual outputs should be seen as supporting tools that improve the transparency of automated systems [58]. Furthermore, as mentioned in [61], Grad-CAM++ by itself is unable to sufficiently clarify the Transformer's decision-making procedure because the hybrid model incorporates a ViT branch. Additional insights into which image patches affect the CLS-token depiction are provided by Transformer-specific techniques like Attention Rollout, Attention Flow Transformer Attribution, and LRP-based Transformer explanations. Future research can therefore prioritize a thorough XAI investigation that incorporates both CNN-side (Grad-CAM++) and ViT-side (Attention Rollout) arguments.

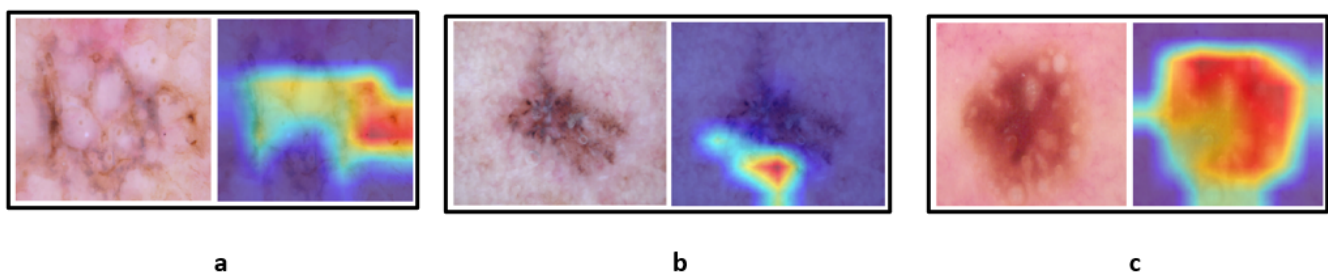


Figure 14. Classes: (a) akeic; (b) bcc; (c) bkl.

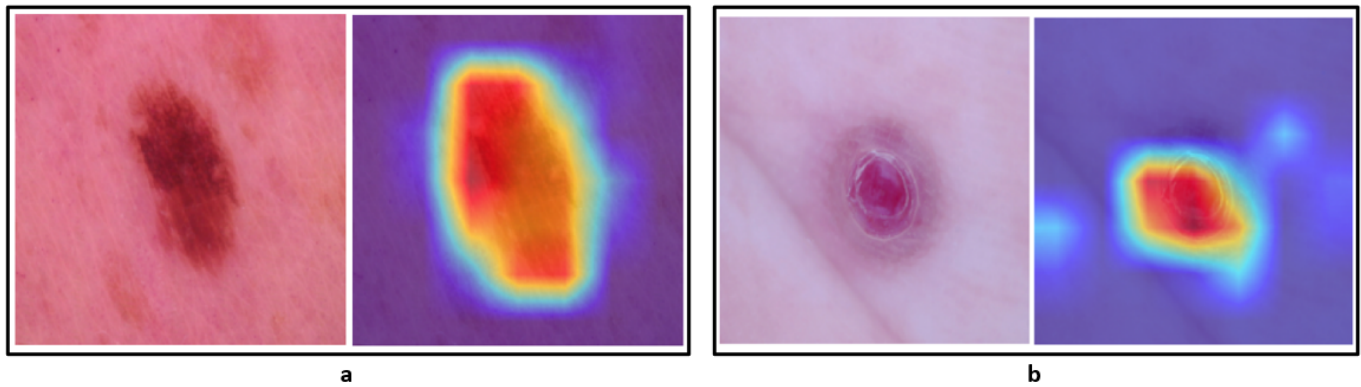


Figure 15. Classes: (a) nv; (b) vasc.

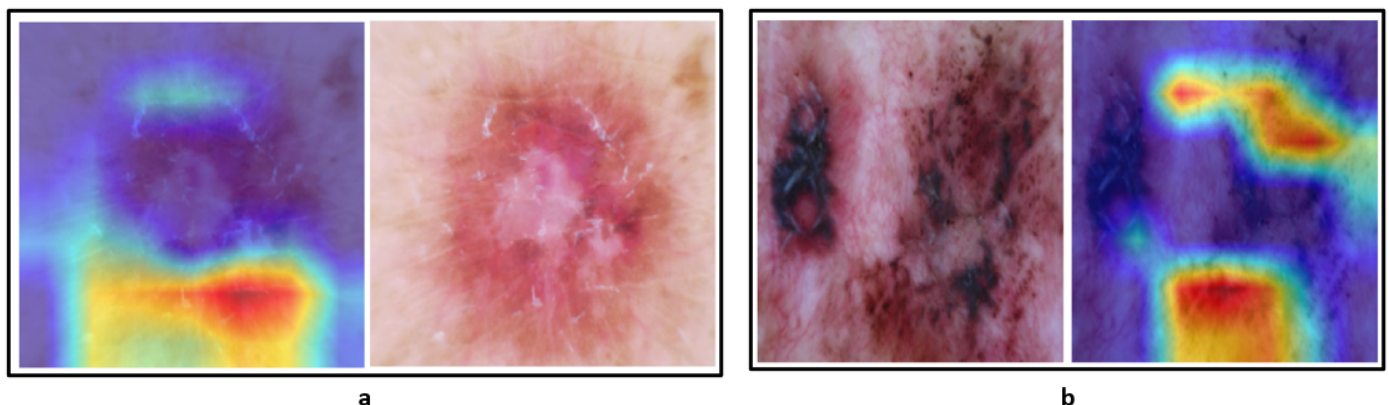


Figure 16. (a) df; (b) mel.

7. Limitations of the Study

The study includes a number of drawbacks that should be noted despite its impressive performance findings. The HAM10000 dataset is the sole basis for the evaluation. Other datasets such as ISIC 2019/2020 [38,62,63] are not used for testing. Models can be validated on multiple datasets to confirm their robustness. Due to their computational cost and high GPU resource requirements for training and inference, the hybrid model and ViT-based architecture are less appropriate for clinical situations with limited resources. Lastly, clinical use of deep learning models is still hampered by their low interpretability, since occasionally they can produce deceptive artifacts or emphasize areas that are not actually related to the underlying disease.

8. Conclusions

Using the HAM10000 dataset, this work suggested a hybrid architecture for skin cancer class classification that incorporates a CNN (ResNet50) and a Vision Transformer (ViT). A thorough comparison between the CNN, ViT, and the suggested hybrid model was also carried out. In terms of overall accuracy and macro-averaged F1-score, the investigational results demonstrate that HybridSkinLes functions better than both the solo CNN and ViT, as well as other existing models studied in the literature. This enhancement demonstrates how well the hybrid model can capture both global contextual linkages and localized spatial elements in dermoscopic pictures. Additionally, the dataset imbalance was partially addressed by the application of class weighting, which led to more comparable results across various lesion types. Nonetheless, disparities in minority class performance continue to exist, which is indicative of the difficulties posed by unbalanced medical datasets. In comparison to the individual models, the hybrid model provides

more reliable classification performance, as further supported by the confusion matrix, ROC curve, and statistical evaluations.

Overall, the results indicate that hybrid CNN–Transformer frameworks, which offer better feature representation and predictive performance, constitute a potential avenue for automated skin lesion categorization. By providing brief visual descriptions for the categorization of skin lesions, this study also enhances the readability of deep learning-based frameworks using Grad-CAM++. In this case, Grad-CAM++ produced consistent, visual mappings that correspond with model decisions and offered explanations. However, there are still issues, including limited generalization because just one dataset was used, significant computing demands, and a lack of external validation. In order to increase clinical applicability and facilitate real-world implementation, further work should concentrate on assessing the model on various datasets, adding clinical information, and improving interpretability.

Author Contributions: Conceptualization, H.G.; methodology, H.G. and M.I.A.; data curation, M.I.A. and H.G.; writing—original draft, H.G. and M.I.A.; draft preparation, H.G.; writing—review and editing, M.I.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The HAM10000 dataset used in this work is publicly available on Kaggle: <https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000> accessed on 10 April 2026.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Shah, A.; Shah, M.; Pandya, A.; Sushra, R.; Sushra, R.; Mehta, M.; Patel, K.; Patel, K. A Comprehensive Study on Skin Cancer Detection Using Artificial Neural Network (ANN) and Convolutional Neural Network (CNN). *Clin. eHealth* **2023**, *6*, 76–84. <https://doi.org/10.1016/j.ceh.2023.08.002>.
- Richard, M.A.; Paul, C.; Nijsten, T.; Gisondi, P.; Salavastu, C.; Taieb, C.; Trakatelli, M.; Puig, L.; Stratigos, A.; EADV Burden of Skin Diseases Project Team. Prevalence of Most Common Skin Diseases in Europe: A Population-Based Study. *J. Eur. Acad. Dermatol. Venereol.* **2022**, *36*, 1088–1096. <https://doi.org/10.1111/jdv.18050>.
- Wolkenstein, P.; Grob, J.-J.; Bastuji-Garin, S.; Ruzsyczynski, S.; Roujeau, J.-C.; Revuz, J. French People and Skin Diseases: Results of a Survey Using a Representative Sample. *Arch. Dermatol.* **2003**, *139*, 1614–1619. <https://doi.org/10.1001/archderm.139.12.1614>.
- Wang, S.Q.; Setlow, R.; Berwick, M.; Polsky, D.; Marghoob, A.A.; Kopf, A.W.; Bart, R.S. Ultraviolet A and Melanoma: A Review. *J. Am. Acad. Dermatol.* **2001**, *44*, 837–846. <https://doi.org/10.1067/mjd.2001.114594>.
- Bickers, D.R.; Lim, H.W.; Margolis, D.; Weinstock, M.A.; Goodman, C.; Faulkner, E.; Gould, C.; Gemmen, E.; Dall, T. The Burden of Skin Diseases: 2004: A Joint Project of the American Academy of Dermatology Association and the Society for Investigative Dermatology. *J. Am. Acad. Dermatol.* **2006**, *55*, 490–500. <https://doi.org/10.1016/j.jaad.2006.05.048>.
- D’Arino, A.; Caputo, S.; Eibenschutz, L.; Piemonte, P.; Buccini, P.; Frascione, P.; Bellei, B. Skin Cancer Microenvironment: What We Can Learn from Skin Aging? *Int. J. Mol. Sci.* **2023**, *24*, 14043. <https://doi.org/10.3390/ijms241814043>.
- Arnold, M.; Singh, D.; Laversanne, M.; Vignat, J.; Vaccarella, S.; Meheus, F.; Cust, A.E.; de Vries, E.; Whiteman, D.C.; Bray, F. Global Burden of Cutaneous Melanoma in 2020 and Projections to 2040. *JAMA Dermatol.* **2022**, *158*, 495–503. <https://doi.org/10.1001/jamadermatol.2022.0160>.
- Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer Statistics, 2019. *CA Cancer J. Clin.* **2019**, *69*, 7–34. <https://doi.org/10.3322/caac.21551>.
- Zunair, H.; Ben Hamza, A. Melanoma Detection Using Adversarial Training and Deep Transfer Learning. *Phys. Med. Biol.* **2020**, *65*, 135005. <https://doi.org/10.1088/1361-6560/ab86d3>.
- Ali, K.; Shaikh, Z.A.; Khan, A.A.; Laghari, A.A. Multiclass Skin Cancer Classification Using EfficientNets—A First Step towards Preventing Skin Cancer. *Neurosci. Inform.* **2022**, *2*, 100034. <https://doi.org/10.1016/j.neuri.2021.100034>.

11. Han, S.S.; Kim, Y.J.; Moon, I.J.; Jung, J.M.; Lee, M.Y.; Lee, W.J.; Won, C.H.; Lee, M.W.; Kim, S.H.; Navarrete-Dechent, C.; et al. Evaluation of Artificial Intelligence-Assisted Diagnosis of Skin Neoplasms: A Single-Center, Paralleled, Unmasked, Randomized Controlled Trial. *J. Investig. Dermatol.* **2022**, *142*, 2353–2362.e2. <https://doi.org/10.1016/j.jid.2022.02.003>.
12. Combalia, M.; Codella, N.; Rotemberg, V.; Carrera, C.; Dusza, S.; Gutman, D.; Helba, B.; Kittler, H.; Kurtansky, N.R.; Liopyris, K.; et al. Validation of Artificial Intelligence Prediction Models for Skin Cancer Diagnosis Using Dermoscopy Images: The 2019 International Skin Imaging Collaboration Grand Challenge. *Lancet Digit. Health* **2022**, *4*, e330–e339. [https://doi.org/10.1016/S2589-7500\(22\)00021-8](https://doi.org/10.1016/S2589-7500(22)00021-8).
13. Borys, K.; Schmitt, Y.A.; Nauta, M.; Seifert, C.; Krämer, N.; Friedrich, C.M.; Nensa, F. Explainable AI in Medical Imaging: An Overview for Clinical Practitioners—Beyond Saliency-Based XAI Approaches. *Eur. J. Radiol.* **2023**, *162*, 110786. <https://doi.org/10.1016/J.EJRAD.2023.110786>.
14. Marchetti, M.A.; Cowen, E.A.; Kurtansky, N.R.; Weber, J.; Dauscher, M.; DeFazio, J.; Deng, L.; Dusza, S.W.; Haliasos, H.; Halpern, A.C.; et al. Prospective Validation of Dermoscopy-Based Open-Source Artificial Intelligence for Melanoma Diagnosis (PROVE-AI Study). *NPJ Digit. Med.* **2023**, *6*, 127. <https://doi.org/10.1038/s41746-023-00872-1>.
15. Escalé-Besa, A.; Yélamos, O.; Vidal-Alaball, J.; Fuster-Casanovas, A.; Miró Catalina, Q.; Börve, A.; Ander-Egg Aguilar, R.; Fustà-Novell, X.; Cubiró, X.; Rafat, M.E.; et al. Exploring the Potential of Artificial Intelligence in Improving Skin Lesion Diagnosis in Primary Care. *Sci. Rep.* **2023**, *13*, 4293. <https://doi.org/10.1038/s41598-023-31340-1>.
16. Olayah, F.; Senan, E.M.; Ahmed, I.A.; Awaji, B. AI Techniques of Dermoscopy Image Analysis for the Early Detection of Skin Lesions Based on Combined CNN Features. *Diagnostics* **2023**, *13*, 1314. <https://doi.org/10.3390/diagnostics13071314>.
17. Das, K.; Cockerell, C.J.; Patil, A.; Pietkiewicz, P.; Giulini, M.; Grabbe, S.; Goldust, M. Machine Learning and Its Application in Skin Cancer. *Int. J. Environ. Res. Public Health* **2021**, *18*, 13409.
18. Jütte, L.; Wang, N.; Steven, M.; Roth, B. Perspectives for Generative AI-Assisted Art Therapy for Melanoma Patients. *AI* **2024**, *5*, 1648–1669.
19. Saeed, M.A.; Afify, Y.M.; Badr, N.L.; Helal, N.A. Multimodal Deep Learning Ensemble Framework for Skin Cancer Detection. *Sci. Rep.* **2025**, *15*, 45660. <https://doi.org/10.1038/s41598-025-30534-z>.
20. Kumar, B.A.; Misra, N.K. Colon Cancer Classification and Detection by Novel CMNV2 Model and Methods of Deep Learning. *Neural Comput. Appl.* **2024**, *37*, 25–41. <https://doi.org/10.1007/s00521-024-10563-x>.
21. Hasan, K.; Ahamad, A.; Yap, C.H.; Yang, G. A Survey, Review, and Future Trends of Skin Lesion Segmentation and Classification. *Comput. Biol. Med.* **2023**, *155*, 106624. <https://doi.org/10.1016/j.combiomed.2023.106624>.
22. Deng, Z.; Yang, Y.; Suzuki, K. Federated Active Learning Framework for Efficient Annotation Strategy in Skin-Lesion Classification. *J. Investig. Dermatol.* **2025**, *145*, 303–311. <https://doi.org/10.1016/j.jid.2024.05.023>.
23. Brancaccio, G.; Balato, A.; Malvey, J.; Puig, S.; Argenziano, G.; Kittler, H. Artificial Intelligence in Skin Cancer Diagnosis: A Reality Check. *J. Investig. Dermatol.* **2024**, *144*, 492–499. <https://doi.org/10.1016/j.jid.2023.10.004>.
24. Ahmed, S.A.A.; Yanikoğlu, B.; Göksu, Ö.; Aptoula, E. Skin Lesion Classification With Deep CNN Ensembles. In Proceedings of the 2020 28th Signal Processing and Communications Applications Conference (SIU), Gaziantep, Turkey, 5–7 October 2020; pp. 1–4.
25. Hasan, K.; Elahi, T.E.; Alam, A.; Jawad, T.; Martí, R. DermoExpert: Skin Lesion Classification Using a Hybrid Convolutional Neural Network through Segmentation, Transfer Learning, and Augmentation. *Inform. Med. Unlocked* **2022**, *28*, 100819. <https://doi.org/10.1016/j.imu.2021.100819>.
26. Jinnai, S.; Yamazaki, N.; Hirano, Y.; Sugawara, Y.; Ohe, Y.; Hamamoto, R. The Development of a Skin Cancer Classification System for Pigmented Skin Lesions Using Deep Learning. *Biomolecules* **2020**, *10*, 1123. <https://doi.org/10.3390/biom10081123>.
27. Thomas, S.M.; Lefevre, J.G.; Baxter, G.; Hamilton, N.A. Interpretable Deep Learning Systems for Multi-Class segmentation and Classification of Non-Melanoma Skin Cancer. *Med. Image Anal.* **2021**, *68*, 101915.
28. Jojoa Acosta, M.F.; Caballero Tovar Liesle Yail and Garcia-Zapirain, M.B.; Percybrooks, W.S. Melanoma Diagnosis Using Deep Learning Techniques on dermoscopic Images. *BMC Med. Imaging* **2021**, *21*, 6.
29. Sevli, O. A Deep Convolutional Neural Network-Based Pigmented Skin Lesion Classification Application and Experts Evaluation. *Neural Comput. Appl.* **2021**, *33*, 12039–12050. <https://doi.org/10.1007/s00521-021-05929-4>.
30. Adablanu, S.; Barman, U.; Das, D.; Dweh, T.J. Transforming Skin Cancer Detection With AI-Based Convolutional and Transformer Models. *iRADIOLOGY* **2026**, *4*, 51–62. <https://doi.org/10.1002/ird3.70058>.
31. Xin, C.; Liu, Z.; Zhao, K.; Miao, L.; Ma, Y.; Zhu, X.; Zhou, Q.; Wang, S.; Li, L.; Yang, F.; et al. An Improved Transformer Network for Skin Cancer Classification. *Comput. Biol. Med.* **2022**, *149*, 105939. <https://doi.org/10.1016/j.combiomed.2022.105939>.

32. Ayas, S. Multiclass Skin Lesion Classification in Dermoscopic Images Using Swin Transformer Model. *Neural Comput. Appl.* **2022**, *35*, 6713–6722. <https://doi.org/10.1007/s00521-022-08053-z>.
33. Cai, Y.; Liu, Y.; Li, J.; Xiong, Q. CDMT-Unet: A Dual-Frequency Cross-Fusion Model for Skin Cancer Segmentation. *Biomed. Signal Process. Control* **2026**, *113*, 108880. <https://doi.org/10.1016/j.bspc.2025.108880>.
34. Halawani, H.T.; Senan, E.M.; Asiri, Y.; Abunadi, I.; Mashraqi, A.M.; Alshari, E.A. Enhanced Early Skin Cancer Detection through Fusion of Vision Transformer and CNN Features Using Hybrid Attention of EViT-Dens169. *Sci. Rep.* **2025**, *15*, 34776. <https://doi.org/10.1038/s41598-025-18570-1>.
35. ThangaPurni, J.S.; Braveen, M. Unified ARP-ViT-CNN System: Hybrid Deep Learning Approach for Segmenting and Classifying Multiple Skin Cancer Lesions. *Array* **2025**, *28*, 100515. <https://doi.org/10.1016/j.array.2025.100515>.
36. Agarwal, S.; Mahto, A.K. Skin Cancer Classification: Hybrid CNN-Transformer models with KAN-Based Fusion. **2025**, *arXiv*, arXiv:2508.12484.
37. Vidales-Basurto, C.A.; Rivera, M. A Deep Learning Framework for Skin Lesion Classification Using ResNet50 and ELM. *Comput. Methods Programs Biomed. Update* **2026**, *9*, 100253. <https://doi.org/10.1016/J.CMPBUP.2026.100253>.
38. Tschandl, P.; Rosendahl, C.; Kittler, H. The HAM10000 Dataset, a Large Collection of Multi-Source Images of Common Pigmented Skin Lesions. *Sci. Data* **2018**, *5*, 180161.
39. Manzari, O.N.; Ahmadabadi, H.; Kashiani, H.; Shokouhi, S.B.; Ayatollahi, A. MedViT: A Robust Vision Transformer for Generalized Medical Image Classification. *Comput. Biol. Med.* **2023**, *157*, 106791. <https://doi.org/10.1016/j.compbiomed.2023.106791>.
40. Nanni, L.; Loreggia, A.; Barcellona, L.; Ghidoni, S. Building Ensemble of Deep Networks: Convolutional Networks and Transformers. *IEEE Access* **2023**, *11*, 124962–124974. <https://doi.org/10.1109/ACCESS.2023.3330442>.
41. Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; Ji, X.; Zhang, Y. TransMIL: Transformer Based Correlated Multiple Instance Learning for Whole Slide Image Classification. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 2136–2147.
42. Medam, S.; Mukkara, A.; Gostu, S.; Medagam, N. Early Non-Invasive Skin Cancer Diagnosis Using ResNet-50 Model. In Proceedings of the 2025 International Conference on Computational Robotics, Testing and Engineering Evaluation (ICCRTEE), Virudhunagar, India, 28–30 May 2025; pp. 1–6.
43. Baker, O.; Ziran, Z.; Mecella, M.; Subaramaniam, K. AI-Driven Melanoma Detection in New Zealand: A ResNet50-Based Approach. *Inform. Med. Unlocked* **2025**, *58*, 101697. <https://doi.org/10.1016/j.imu.2025.101697>.
44. Zhou, J.; Wu, Z.; Jiang, Z.; Huang, K.; Guo, K.; Zhao, S. Background Selection Schema on Deep Learning-Based Classification of Dermatological Disease. *Comput. Biol. Med.* **2022**, *149*, 105966. <https://doi.org/10.1016/j.compbiomed.2022.105966>.
45. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A Survey on Vision Transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 87–110. <https://doi.org/10.1109/TPAMI.2022.3152247>.
46. Kılıç, Ş.; Doğan, Y. A Pruned and Parameter-Efficient Xception Framework for Skin Cancer Classification. *PLoS ONE* **2026**, *21*, e0341227.
47. Unnisa, Z.; Tariq, A.; Sarwar, N.; Serhani, M.A.; Trabelsi, Z. Impact of Fine-Tuning Parameters of Convolutional Neural Network for Skin Cancer Detection. *Sci. Rep.* **2025**, *15*, 14779. <https://doi.org/10.1038/s41598-025-99529-0>.
48. Pacal, I.; Ozdemir, B.; Zeynalov, J.; Gasimov, H.; Pacal, N. A novel CNN-ViT-based deep learning model for early skin cancer diagnosis. *Biomed. Signal Process. Control* **2025**, *104*, 107627. <https://doi.org/10.1016/j.bspc.2025.107627>.
49. Yang, G.; Luo, S.; Greer, P. A Novel Vision Transformer Model for Skin Cancer Classification. *Neural Process. Lett.* **2023**, *55*, 9335–9351. <https://doi.org/10.1007/s11063-023-11204-5>.
50. George, J.; Rao, A. A Three-Tier BERT Based Transformer Framework for Detecting and Classifying Skin Cancer with HSCGS Algorithm. *Multimed. Tools Appl.* **2023**, *83*, 51441–51467. <https://doi.org/10.1007/s11042-023-17590-1>.
51. Najjar, F.H.; Khudhair, Z.N.; Mohamed, F.; Rahim, M.S.M.; Chan, V.S.; Ali, A.H. Transformer-Aided Skin Cancer Classification Using VGG19-Based Feature Encoding. *Sci. Rep.* **2025**, *15*, 40204. <https://doi.org/10.1038/s41598-025-24081-w>.
52. Aboulmira, A.; Hrimch, H.; Lachgar, M.; Hanine, M.; Garcia, C.O.; Mezquita, G.M.; Ashraf, I. Hybrid Model with Wavelet Decomposition and EfficientNet for accurate Skin Cancer Classification. *J. Cancer* **2025**, *16*, 506–520.
53. Al Shafi, A.; Muntakim, A.; Shill, P.C.; Zannat, R.; Al-Amin, A. Skin Lesion Classification Using a Soft Voting Ensemble of Convolutional Neural Networks. In Proceedings of the 2025 International Conference on Electrical, Computer and Communication Engineering, ECCE 2025, Chittagong, Bangladesh, 13–15 February 2025; Institute of Electrical and Electronics Engineers Inc.: New York, NY, USA, 2025.

54. Al Sakib, A.; Swapno, S.M.R.; Ahamed, F.; Mohiuddin, A.B.; Bhuiyan, M.I.H.; Khan, S.; Khushbu, K.G.; Haque, R.; Alahmadi, T.J.; Moni, M.A. Explainable AI-Driven Hybrid Deep Learning Framework for Accurate Skin Cancer Diagnosis. *Digit. Health* **2026**, *12*, 1–35. <https://doi.org/10.1177/20552076261438923>.
55. Kreouzi, M.; Theodorakis, N.; Feretzakis, G.; Paxinou, E.; Sakagianni, A.; Kalles, D.; Anastasiou, A.; Verykios, V.S.; Nikolaou, M. Deep Learning for Melanoma Detection: A Deep Learning approach to Differentiating Malignant Melanoma from Benign Melanocytic. *Cancers* **2024**, *17*, 28.
56. Lu, Y.; Wang, M.; Zhang, G. A New Revised Version of McNemar's Test for Paired Binary Data. *Commun. Stat. Theory Methods* **2017**, *46*, 10010–10024. <https://doi.org/10.1080/03610926.2016.1228962>.
57. Jafari, M.; Ansari-Pour, N. Why, When and How to Adjust Your P Values? *Cell J.* **2018**, *20*, 604. <https://doi.org/10.22074/CELLJ.2019.5992>.
58. Fiaz, M.; Shoaib Khan, M.B.; Khan, A.H.; Bilal, A.; Abdullah, M.; Darem, A.A.; Sarwar, R. An Explainable Hybrid Deep Learning Framework for Precise Skin Lesion Segmentation and Multi-Class Classification. *Front. Med.* **2025**, *12*, 1681542. <https://doi.org/10.3389/fmed.2025.1681542>.
59. Bhati, D.; Neha, F.; Amiruzzaman, M. A Survey on Explainable Artificial Intelligence (XAI) Techniques for Visualizing Deep Learning Models in Medical Imaging. *J. Imaging* **2024**, *10*, 239. <https://doi.org/10.3390/jimaging10100239>.
60. Gamage, L.; Isuranga, U.; Meedeniya, D.; De Silva, S.; Yogarajah, P. Melanoma Skin Cancer Identification with Explainability Utilizing Mask Guided Technique. *Electronics* **2024**, *13*, 680. <https://doi.org/10.3390/electronics13040680>.
61. Neha, F.N.U.; Bansal, A. Understanding the Architecture of Vision Transformer and Its Variants: A Review. In Proceedings of the 1st International Conference on Innovative Engineering Sciences and Technological Research, ICIESTR, Muscat, Oman, 14–15 May 2024. <https://doi.org/10.1109/ICIESTR60916.2024.10798341>.
62. Codella, N.C.F.; Gutman, D.; Celebi, M.E.; Helba, B.; Marchetti, M.A.; Dusza, S.W.; Kalloo, A.; Liopyris, K.; Mishra, N.; Kittler, H.; et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). *arXiv* **2017**, arXiv:1710.05006.
63. Hernández-Pérez, C.; Combalia, M.; Podlipnik, S.; Codella, N.C.F.; Rotemberg, V.; Halpern, A.C.; Reiter, O.; Carrera, C.; Barreiro, A.; Helba, B.; Puig, S.; Vilaplana, V.; Malvehy, J. BCN20000: Dermoscopic lesions in the wild. *Sci. Data* **2024**, *11*, 641. <https://doi.org/10.1038/s41597-024-03387-w>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.