Pertanika J. Sci. & Technol. 34 (2): 625 - 646 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Skin Cancer Classification from Dermoscopic Images based on Convolutional Neural Network and Vision Transformer

Gan Mengyun[1], Hazlina Hamdan[1]*, Mahmud Dwi Sulistiyo[2], and Abdulrahman Dira Khalaf [1,3]

[1]Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia
[2]School of Computing, Telkom University, Jl. Telekomunikasi No. 1, Terusan Buah Batu, 40257 Bandung, Jawa Barat, Indonesia
[3]Computer Centre Department, University of Fallujah, 31002 Fallujah, Anbar, Iraq

ABSTRACT

Skin cancer poses a vast worldwide health issue because of its fast progression and high mortality rate. Early diagnosis and precise prognosis are important for the treatment of these patients. In this paper, we propose an enhanced hybrid deep learning model composed of a combination of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for the dermoscopic image classification task. The designed architecture adopts ResNet50, EfficientNet, and ViT as strong feature extractors with an innovative attention module, MambaATT, for precise modelling and utilisation of global context information. Methods for statistical growth (rotation, flips, and scaling) have been tested to improve the model's generalisation and its variety, even if the dataset is short. When evaluating the entire dataset, we found that our proposed version CNN_ViT_MambaATT reaches an advanced-class accuracy of 92%, outperforming the traditional methods for individual networks. The findings highlight the power of combining strong CNN and VIT models for both mechanisms of interest and information growth strategies, offering a robust and accurate approach to early skin cancer analysis.

Keywords: Convolutional neural networks, data augmentation, deep learning, EfficientNet, MambaATT Model, ResNet50, skin cancer, vision transformer

ARTICLE INFO

Article history:
Received: 21 October 2025
Accepted: 26 December 2025
Published: 01 April 2026

DOI: https://doi.org/10.47836/pjst.34.2.02

E-mail addresses:
gsgs84617@gmail.com (Gan Mengyun)
hazlina@upm.edu.my (Hazlina Hamdan)
mahmuddwis@telkomuniversity.ac.id (Mahmud Dwi Sulistiyo)
adk1973@uofallujah.edu.iq (Abdulrahman Dira Khalaf)
* Corresponding author

INTRODUCTION

Cancer is a medical condition consisting of deviant proliferation, division and distribution of cells throughout the human body by way of the lymphatic system and

e-ISSN: 2231-8526

blood, subsequently destroying healthy tissues. Skin cancer is one of the most frequent forms of cancer worldwide. It is caused by the overgrowth of skin cells, usually induced by excessive contact with UV light or carcinogens. Quickened development of modern industry and chemical industry Results to pollution and disaster for environment deployed Environmental factors such as industrial pollution and ozone depletion have led to a global rise in skin cancer occurrence. For 2025, it is expected that there will be approximately 104,960 new cases of melanoma in the United States, with around 60 550 in men and roughly 44,410 in women. Similarly, the projected number of deaths as 5,470 and 2,960 for males and females, respectively (Siegel et al., 2025). New cases in the USA, according to projections by sex, are shown in Figure 1. These numbers highlight the critical need for early and precise diagnostics.

The most lethal are the rapidly advancing, high-mortality tumours. In particular, the first identity complements the survival rate and emphasises that precisely adequate clinical equipment is necessary. The symptoms of skin cancer in the early stage are like those of a non-cancerous mole. "Low inter-class variance and high intra-class variance" in skin lesions is an extraordinary feature that brings a great challenge to dermatology specialists. Diagnosing such diseases by purely visual characteristics assisted by nonspecific histopathological biopsy could be misdiagnosed and have loss of diagnosis, and this would make patients lose their best treatment time. To increase the accuracy of skin

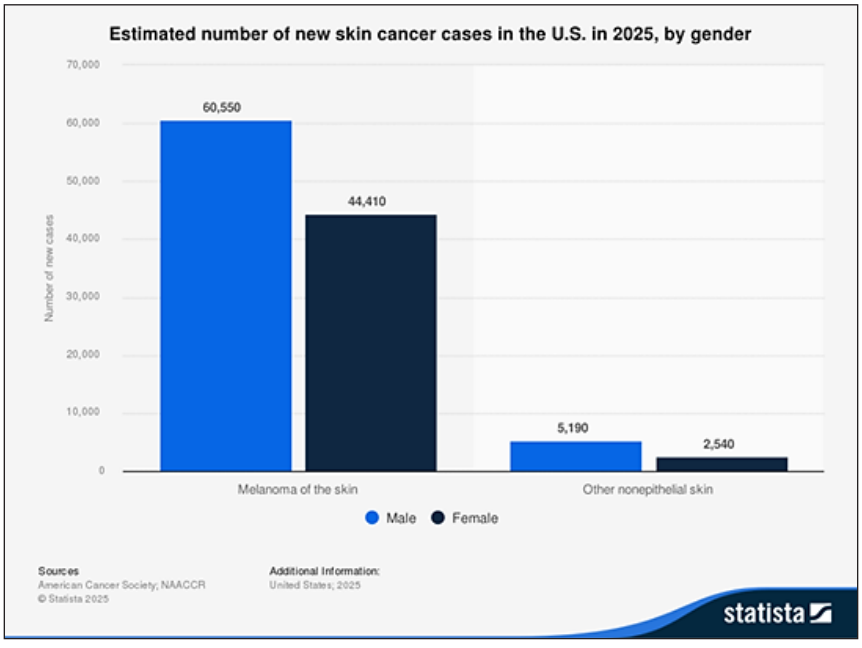

Figure 1. Estimated number of new skin cancer cases in the U.S. in 2025, by gender (Siegel et al., 2025)

disease diagnosis, the dermoscopy diagnostic technique has been developed and has been working well in clinical analysis (Kittler et al., 2002). Dermoscopy is a surface imaging method that utilises optical imaging for deeper skin lesions resulting in appearance of most affected areas.

In comparison to a single visual inspection, dermoscopy allows experts to diagnose and evaluate lesions under the direct observation with high resolution images, thus offering more accurate treatment plans for patients and leading to higher survival of skin cancer patients. Nevertheless, because the structure of dermoscopic images is so complex, its accuracy largely relies on physicians' expertise. The challenge is in the variability of the skin lesions which makes automatic classification very hard. This is primarily due to features common among diverse skin lesions can give rise to misdiagnoses. Second, it changes within the same category partially overlap with those of the lesion subtype (Goyal et al., 2020). A more accurate, rapid and objective method for automatic classification of skin cancer is therefore of great importance for the patients as well as medical professionals.

Deep learning has advanced and skin cancer classification is no exception including convolutional neural networks (CNNs) and vision transformers (ViTs). CNNs can be recognised by utilising local spatial features information in dermoscopic images, while ViTs use a self-attention mechanism. This work critically investigates the development of CNN and ViT architectures by demonstrating impactful hybrid models that harness their individual advantages. These hybrid processes have improved lesion detection and classification, providing a potential solution to current drawbacks of dermatology practices. Skin cancer in particularly melanoma, remains an important global health challenge due to its aggressive behaviour and associated fatal outcome when diagnosed late. Data-driven diagnostic systems have ditched a rule-based framework to enhance clinical performance and robustness by employing deep learning methods. CNN is the conventional backbone for dermoscopic image processing. The CNN structure is partitioned into local spatial patches, and it proves very effective in the image-based skin ulcer classification. As CNNs can capture hierarchical features from image data, they have been significant in the field of medical image analysis. Baykal Kablan and Ayas (2024) proposed an automatic clothing recognition function using a convolutional model with 97.1% accuracy on the ISIC 2019 dataset. It is a paradigm shift to treat images as sequential data instead of CNNs, which enables the recovery of coherence on the global context.

Recently, ViT was introduced in medical imaging, and it has attracted significant attention for its ability to capture global context through self-attention. A combining of CNN, ViT, and Xception introduced model achieving 95.46% precision and 96.74% accuracy on the HAM10000 dataset. The model effectively captured both local details and global contexts, proving robust in complex classification tasks (Ali et al., 2025). Reis and Turk (2024) suggested the MABSCNET model, a hybrid VIT architecture, which

received 100% accuracy on the IC 2018 data set with significant performance (92.74%) with a combination of CNN and ViT architecture. Arshed et al. (2023) reported the efficacy of ViT architecture over traditional CNN models, achieving 92.14% accuracy in multi-class skin cancer classification. A hybrid CNN-transformer model with focal losses to address class imbalance. This model effectively used CNN for early local functional extraction and a transformer model for global meaningful interpretation, and significantly improved classification performance in ISIC 2018 (Nie et al., 2023). Yang et al. (2025) presented a new attention-based approach at multiple levels, increasing the exact speed by 95.05%. Abbas et al. (2023) proposed Assist-Dermo, a lightweight transformer-based CNN hybrid achieving 95.6% accuracy through optimised architecture and preprocessing strategies.

One of the challenges of early detection is that malignant and benign tumours are often indistinguishable from one another. Early detection is vital; if the diagnosis is made at early stage, cure rate is more than 95% while it plunges to approximately 14% in advanced stages (Khan et al., 2021). Their morphological resemblance, size and pigmentation make them frequently being misdiagnosed by visual observation and histological diagnosis only. This limitation might delay treatment and contribute to lower overall survival. Dermoscopy has been widely used to improve the accuracy of diagnosis. This non-invasive imaging tool improves the visualisation of subsurface skin structures and treatment planning (Kittler et al., 2002). Nevertheless, accuracy of dermoscopic diagnosis is largely dependent on the experience of the physician while interobserver variation remains an issue. Because of skin lesions' intrinsic variability, it is challenging to classify skin lesions automatically. ViTs interpolate between the model and classic provisions in case of an optimal amount of regularisation and capacity (Flosdorf et al., 2024). This way skin cancer will be automatically classified with the patient and doctor faster, more accurately, and unbiasedly.

An important aspect of the investigations that have been performed on DL models for skin cancer classification is to compare the metrics such as accuracy, sensitivity, specificity, and F1 score. Research reported that the hybrid models outperform CNN or VIT alone. Despite some progress, various challenges bias, lesion-awareness uncertainty, real-time deployment, and lack of diversity or clarity remain. The fusion of CNNs and ViTs has boosted the skin cancer classification task by a considerable margin. Recent surveys Meedeniya et al. (2024) highlight the absence of integrated architectures combining CNNs, Vision Transformers (ViTs), and state-space models in medical imaging. In this paper, to cope with these issues in automatic diagnosis of skin cancer, we propose the combination of convolutional and transformer models. Despite significant advancements, there are still some important missing links in combining the local and global feature learning into effective and interpretable networks. Three central questions guide the research:

1. Can a hybrid CNN–ViT–MambaATT architecture outperform individual CNN and ViT models for dermoscopic image classification?

2. Does the MambaATT state-space mechanism improve discriminative feature extraction compared to traditional attention?
3. How does multi-branch feature fusion influence performance on limited medical datasets?

Recent state-of-the-art methods for skin cancer image classification have shown good to mixed levels of performance. Huang et al. (2024) used an ensemble-based attention CNN model, and their accuracy was 87.8% while Naz et al. (2025) achieved an accuracy of 86.5% using federated VGG19-XAI architecture. In another study, Lea (2025) used topological data analysis along with machine learning and presented 91.25% accuracy on a binary dataset of a similar kind. Nevertheless, the performance accuracy is still low. This study's novelty lies in combining the strengths of ResNet50, EfficientNet, and ViT via tensor-based feature fusion, and in refining the resulting features with the lightweight MambaATT module. The proposed approach captures both spatial and contextual dependencies with fewer parameters, improving accuracy and interpretability.

The objective of our study is to construct an accurate diagnostic model for differentiating dermoscopic tumour images as benign or malignant, with these three contributions:

1. Applying data augmentation to enhance generalisation with small dataset.
2. A composite model consisting of ResNet50 (He et al., 2016), EfficientNet (Tan & Le, 2019), ViT and MambaATT.
3. Evaluation of the proposed model against bench models to show improved classification accuracy.

This paper is organised as follows: Section 2 provides background on the research domain and motivation to work in the field of skin cancer classification: this includes a discussion of current state-of-the-art research, applied methodology and an overview of relevant literature addressing the problem discussed throughout. Section 3 describes the approach in this study and presents information on algorithms and operations. Implementation steps and results of the implementation are presented in Section 4. The reproduction of the base work and the improvement are described and compared. Section 5 provides the conclusion of the paper. Finally, Section 6 discusses limitations and future research.

RELATED WORK

In recent years, deep learning has brought significant improvement in medical imaging and outperformed conventional machine learning for several tasks, including skin cancer classification. In this section, we briefly describe the main building blocks of the deep learning image classification landscape. The emphasis is on convolutional neural network (CNN) architectures and their development from early ideas to the current state of the

art. The power of transfer learning and ensemble training is explored, emphasising how they can be utilised across a range of tasks. Vision Transformer and Vision Mamba are particularly highlighted, together with discussing techniques that affect model generalisation and feature extraction.

Since the introduction of AlexNet in 2012, deep neural networks have become the workhorse for computer vision applications, especially dermatological image diagnosis (Krizhevsky et al., 2017). Unlike traditional methods, where the features are designed manually, deep learning can automatically learn detailed patterns from input data based on convolutional and pooling operations. CNNs are the most common architecture in this area. They effectively model the spatial hierarchy in images using a localised representation and have shown reliability on different classification problems (LeCun & Bengio, 1998). CNN-based models, including AlexNet (Krizhevsky et al., 2017), DenseNet (Mobiny et al., 2019), and ResNet (He et al., 2016) have been widely used in skin cancer research. Some studies have improved the performance of convolutional neural networks for skin lesion classification. Han et al. (2020) included attention mechanisms into ResNet-50, resulting in a sensitivity of 76.8% and a specificity of 90.6%, while Tschandl et al. (2020) documented an accuracy of 80.3% on the HAM10000 dataset utilising ResNet-34. Mahbod et al. (2020) developed a multi-scale fusion CNN achieving 86.2% accuracy on the same dataset. In a different approach, Ali et al. (2022) employed EfficientNet in conjunction with transfer learning and preprocessing techniques, including hair removal, to attain an accuracy of 87.91%. Huang et al. (2021) developed a lightweight deep learning network for the classification of skin cancer, with an accuracy of 85.8%.

Despite this development, CNN's obstacles to capturing global features, researchers use transformer-based architecture, especially the visual transformer (VIT), which uses a self-attention technique to achieve global features (Chen et al., 2020; Dosovitskiy et al., 2021). Unlike CNNs, which need to stack several layers for perceiving the global patterns, Wits might capture relations between widely separated regions in one layer. The main idea of ViT is to cut the input images into non-overlapping patches and transform them into embeddings, which are subsequently fed through stacked transformer layers aided by self-attention. This enables the model to reason about the relative significance of many image regions. Below are some of the advantages that Vision Transformers have over traditional CNNs:

- **Stringer Long-range Information:** The CNNs are sensitive to the input size and objects in large images, thus making it hard for them to capture the global context. Vision Transformer models can capture long-range feature relations in images by self-attention, hence achieving better global information extraction.
- **Scale Invariances:** Traditional CNN-based models use fixed-size input, which makes it difficult to deal with images of other resolutions. Vision Transformer

models that cut images into small patches can better generalize to the input images of different resolutions.

- **Better Interpretability:** The output of Vision Transformer can indicate the importance of patches, which leads to a better understanding of the decision model.

It has been shown in the recent state of the art that hybrid models using either CNN or transformers are successful for skin cancer detection. For example, Pon Selchiya et al. (2023) presented a CNN-ensemble for melanoma classification and Ali et al. (2025) developed xCViT, where the authors combined a CNN, ViT, and Xception model for better explainability. However, these architectures are based only on self-attention operators and do not consider latent state-space modelling. Abbas et al. (2023) studied ensemble hybrid networks, nevertheless, without a systematic way to efficiently model the long-range features. The CNN_ViT_MambaATT differs by employing the Mamba module, which is a selective state-space mechanism designed to permit token mixing at fewer computational costs, allowing for better context memory.

Another study by Ozdemir and Pacal (2025) demonstrated that VIT can outperform CNN on multiclass classification when regularisation and followed training are enhanced. Vit enhances the generalisation by showing that patches indeed matter in the decision. However, some challenges are faced. They are also less biased than CNNs and need large datasets (Reis & Turk, 2024; Tay et al., 2021). Furthermore, the constrained local modelling capacity and overfitting problem within deep layers would also hamper its performance. More recent studies have attempted to loosen these constraints by looking at state space models (SSMs), such as Mamba (Gu & Dao, 2023), whose goal is to allow the modelling of linear time series with a bigger motivational bias. Zhu et al. (2024) proposed a visual-specific SSM with a good trade-off for 2D-selective scanning, capturing both spatial and sequential information. This changing environment indicates the necessity for a coupling design that allows for taking the advantages of CNNs, transformers, and SSMs.

METHODOLOGY

In this section, the architecture of our project is described, with deep learning model architectures CNN_ViT_MambaATT and transfer learning techniques that employ pre-trained ResNet50, ViT, and EfficientNet. The confusion matrix, as well as accuracy, precision, recall, and F1 score, are used to examine the performance of the model.

Features were taken from the penultimate convolutional block of ResNet50, EfficientNet, and from the final transformer encoder output of ViT. These high-level representations encode complementary visual semantics: local patterns in CNNs and global relationships in ViT. The reference and query features were stacked using a tensor stacking operation to form a single representation along a new dimension. The stacked tensor was further fed into MambaATT for learning sparse spatial-temporal dependencies. MambaATT is used as a post-fusion refinement module and does not serve as a substitute

to any CNN or ViT block. It directly manipulates the concatenated feature tensor, improving inter-model dependencies via selective state transition. To counteract the impact of contributions between backbones, a weighted average operation was performed prior to fusion. A Kaggle community dataset based on the ISIC Archive is used for this study. The dataset contains a total of 3,297 dermoscopic images divided into benign and malignant images. The dataset is balanced between two classes Table 1, allowing for a fair skin cancer classification comparison between different approaches.

Table 1
Distribution of training and test sets in the ISIC dataset

Type	Train	Test	Total
Benign	1440	360	1800
Malignant	1197	300	1497
Total	2637	660	3297

The medical images have been pre-processed and scaled to 224×224 pixels. The dataset is split into a training set 80% and a testing set 20%. This method ensures that a significant fraction of the data is used for model training, enabling the model to learn from a large dataset. Concurrently, allocating 20% of the data for testing facilitates a significant assessment of the model's efficacy on unfamiliar data. This method identifies overfitting when the model excels on training data but underperforms on novel data, thereby ensuring the model's generalisability. Figure 2 shows benign and malignant images taken from the dataset.

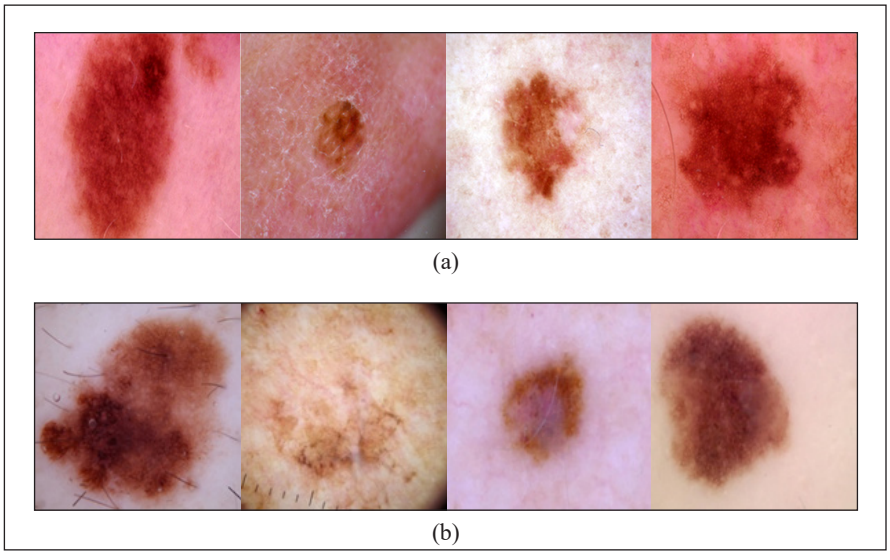

Figure 2. Example of images taken from the ISIC dataset (a-Benign, b-Malignant)

Dataset Preparation

Data augmentation refers to the process of applying a variety of transformation operations on some source images in a small-sized image dataset to create additional training data

that can help enhance the model's generalisation (Mikołajczyk & Grochowski, 2018). Since the dataset size is relatively small, we augment the training data by horizontally and vertically flipping and randomly rotating from 0 to 45 degrees. Figure 3 shows images with different sorts of augmentations applied. By doing so, the dataset is enlarged and enables the trained network to better classify data.

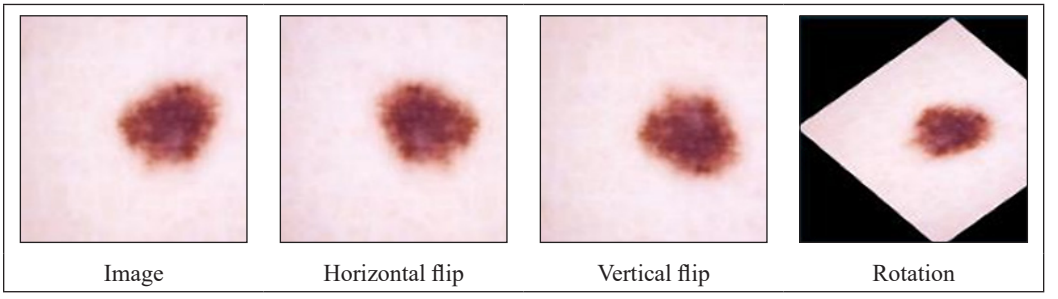

Figure 3. Comparison of the images transformed in different ways

CNN_ViT_MambaATT

Skin cancer image classifiers must use the datasets in a large size, and features extraction from skin image analysis is considered one of the major problems in this application. To solve this, the CNN_ViT_MambaATT module is proposed with a hybrid architecture for image classification on public datasets using transfer and ensemble learning. The CNN_ViT_MambaATT module provides an accurate diagnosis of skin diseases and determination of the malignancy or benignity. Combining CNNs with ViTs exploits the advantages of both models. CNNs are good at extracting local information, while self-attention in ViTs focusses on modelling global interaction information. This combined method will enhance general performance on image classification. On the contrary, CNNs generalise better with less data and achieve higher accuracy compared to ViTs. On the contrary, ViTs are efficient to learn from a small number of images. The general flowchart of the process is illustrated in Figure 4.

Feature Extraction

This study utilises a hybrid feature extraction framework, involving the fusion of 3 already pre-trained deep learning architectures: ResNet50, EfficientNet and Vision Transformer (ViT). These models separately process the input photos and produce high-dimensional feature vectors that capture diverse aspects of the image data. ResNet50 employs residual learning that allows for training of deeper networks and the capture of intricate low-level and mid-level visual patterns. Our EfficientNet introduces a new efficient model scaling method that balances all dimensions of depth/width/resolution and leads to better

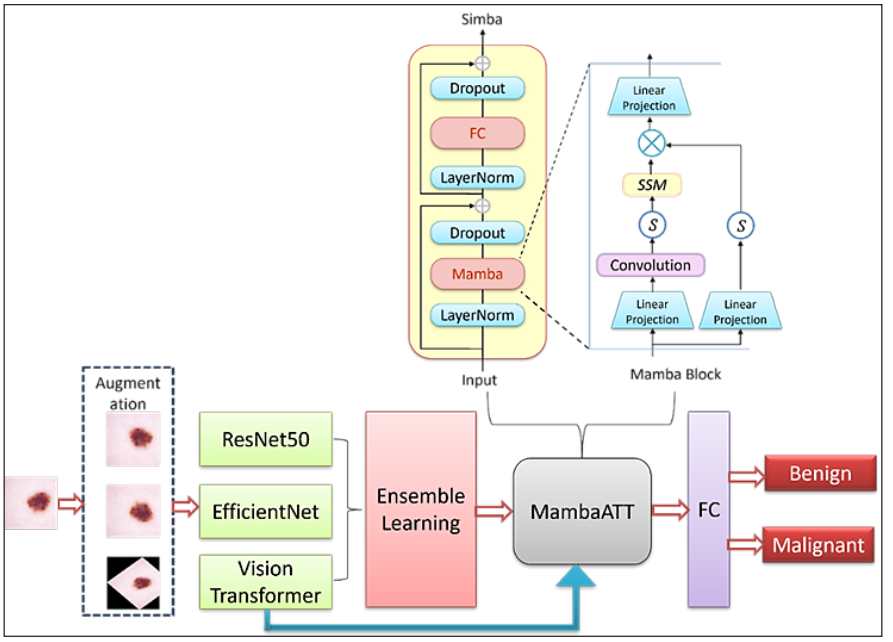

Figure 4. The proposed framework of skin cancer classification

performance. ViT applies self-attention to image patches so that visual structures can be generally recognised. We select the model out of architectural considerations and to have an effective way to benchmark image classification performance. Autonomous processing implies that each model has distinct functionality, leading to diversity and resistance.

In ensemble learning, stack the ResNet50, EfficientNet, and ViT features using tensor stacking. This approach concatenates the feature vectors along one extra tensor dimension, so that subsequent layers can learn from all three models' full spectrum of information. Ensemble learning is essential to facilitate generalisation and predictive power. Architecture combines multiple types to alleviate the shortcomings of individual models while exploiting their advantages. This is especially useful in medical imaging, where lesions may differ greatly in shape, colour, and texture. For example, the ViT may be able to recognise global context while ResNet50 and EfficientNet may have an advantage in detecting localised or hierarchical features.

Attention-based Weighting Mechanism

As illustrated in Figure 5, to reconcile the cooperation of three feature extractors ResNet50, EfficientNet, and ViT, an attention-guided weighting strategy is employed before feature fusion. For both backbones, a feature vector was obtained that was normalised and processed with a learned attention gate. This gate was a pluggable lightweight

MLP with a sigmoid activation function, and it produced an importance score for each feature branch. These coefficients, which expressed the importance of the feature of each model for the given input (i.e., in a local scope), were used to calculate the weighted sum of representations in Eq. 1.

$$F_{\text{fused}} = \sum_{i=1}^{3} \alpha_i \times F_i \quad [1]$$

$$\text{where } \alpha_i = \text{sigmoid}(W_i F_i + b_i) \text{ and } \sum_i \alpha_i = 1$$

Such adaptive weighting enables the model to highlight CNN features for texture-rich regions as well as ViT features for global structural context and leads to dynamic and feature fusion. Architecture allows the model to dynamically focus on texture-based local information extracted with CNNs, while preserving richer and more semantic contextuality represented by ViT, leading to improvements in both representation space richness and classification performance.

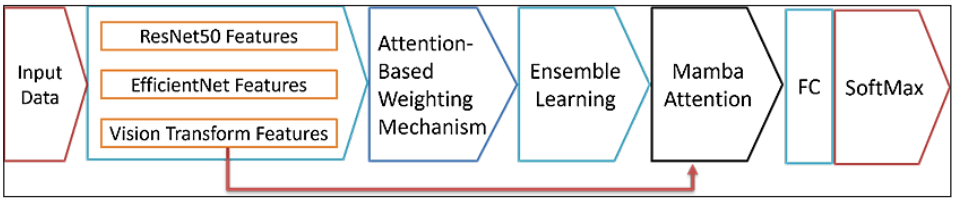

Figure 5. Proposed CNN_ViT_MambaATT framework

MAMBA Attention

We propose MambaATT, an attention mechanism specifically designed to improve the discriminative capacity of combined features. At the expense of creating additional parameters, MambaATT weighs different feature map parts in the pattern that corresponds to their relevance for a classification purpose. The method uses enhancing features that reinforce certain characteristics while suppressing irrelevant details or redundancy. This attention helps the model focusses on the most discriminative regions between benign and malignant tumours, especially for cases with subtle visual clues. MambaATT is intended to work in a distributed manner across the network to decrease the number of parameters of the network while maintaining high precision. That is, since the network can attend to whatever it needs to, it does not require the parameters that would be necessary if all the inputs were being processed. The experimental results show that the proposed extraction strategy based on the ensemble deep learning model and MambaATT modules can

effectively improve classification accuracy and robustness. This method offers a practical and potentially effective downsampling in CADs of dermatology. Figure 6 illustrates the structure of Mamba attention (MambaATT). MambaATT takes the stacked feature tensor and proceeds to process the features, producing an attention-weighted representation. To improve the stability and accelerate the training process, layer normalisation is used to normalise the input among features. A dropout layer, which deactivates a random subset of neurons during training to reduce overfitting, enhances model generalisation. Mamba for token mixing addresses issues of inductive bias and computational complexity. Finally, the attention-weighted features are flattened into a single vector and passed through two fully connected (FC) layers for final classification. Through the steps, the features obtained from MambaATT combine with the feature vectors acquired from the previous ViT model. Finally, these combined features are fed into the FC layer for the result analysis.

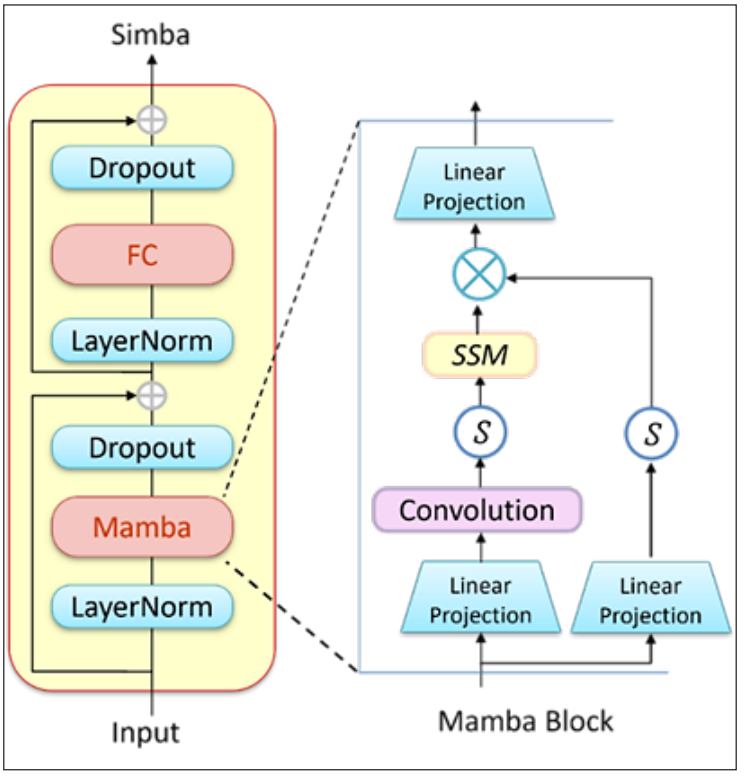

Figure 6. The architecture of MambaATT

Training and Evaluation

The accuracy and loss plots in Figure 7 show that the CNN_ViT_MambaATT model reaches a plateau after only a few epochs, with no sign of overfitting during training.

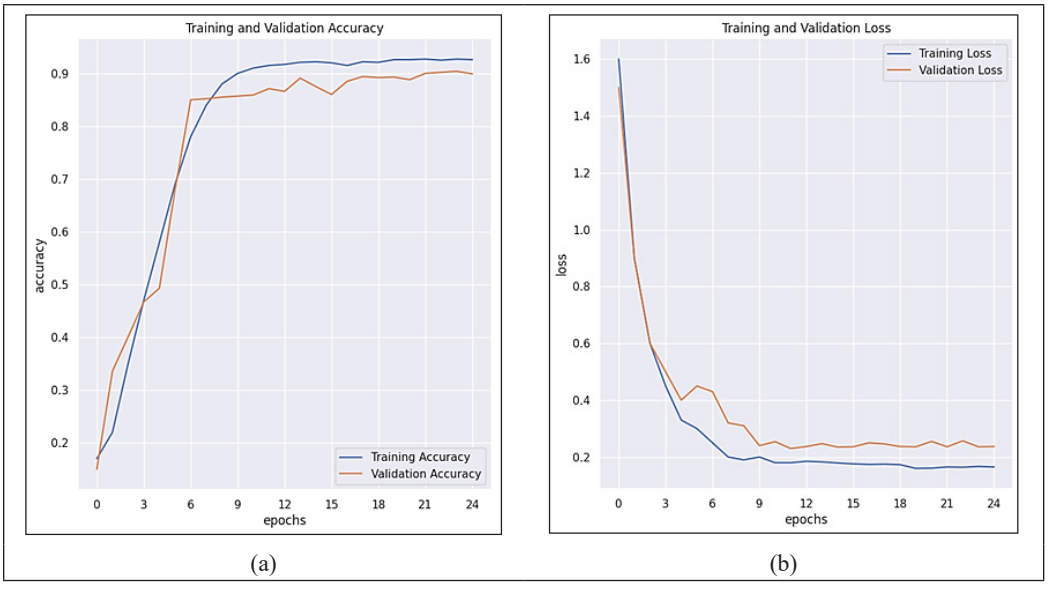

Figure 7. Training and validation (a) accuracy and (b) loss over epochs for CNN_ViT_MambaATT

We employed transfer learning techniques to leverage pre-trained ResNet50, EfficientNet, and ViT models trained on the ImageNet dataset. These models have already acquired valuable hierarchical features from the extensive and diverse ImageNet dataset. Utilising the knowledge acquired during their ImageNet training proves particularly advantageous when handling small datasets. This approach significantly reduces the training time and computational resources required for our research. The only modification that we made was in the last fully connected layer to have 2 output classes. These three models were used as base models for feature extraction. The combined extracted features were processed through "torch. stack," stacking tensors along a new dimension to form their complete feature representation. Then, the custom core module MambaATT was instantiated, and the features output by pre-trained models were concatenated as the input of MambaATT. In MambaATT, features were additionally extracted by its attention mechanism and multi-layer perception to promote the accuracy of image classification.

To assess the power of the feature set learned by CNN_ViT_MambaATT for skin cancer dermoscopic image classification, we compared it in three network architectures. Experiments reused the same datasets and widely used model evaluation metrics to ensure comparability of outcomes. The training employed a scheduling learning rate, starting from 0.001. The learning rate was divided by 10 if the model did not converge within two epochs. The training configuration for all the experiments is shown in Table 2. The MambaATT module was used with a state dimension of 64 an expansion factor was set to 2 and the size of convolution kernel was 3. This process was followed by a dropout rate of 0.1 to alleviate overfitting and the GELU activation function for improved nonlinearity.

Then two MambaATT layers were cascaded to refine the fused features from ResNet50, EfficientNet and ViT. These hyperparameters were determined empirically based on validation performance to balance model complexity and classification accuracy. The evaluation equations are given by:

$$\text{Recall(Sensitivity)} = \frac{\text{TP}}{\text{TP+FN}} \quad [2]$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad [3]$$

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad [4]$$

$$\text{F1/Dice} = \frac{2\text{TP}}{2\text{TP+FP+FN}} \quad [5]$$

$$\text{True Positive Rate (TPR)} = \frac{\text{TP}}{\text{TP+FN}} \quad [6]$$

$$\text{False Positive Rate (FPR)} = \frac{\text{FP}}{\text{FP+FN}} \quad [7]$$

Table 2
Training parameters for each experiment

Parameter Name	Value
Input image dimension	224×224×3
Batch size	64
Learning rate	0.001
Training Epochs	25
Loss function	Cross-Entropy Loss
Optimiser	Adam
State dimension (MambaATT)	64
Expansion factor (MambaATT)	2
Convolution kernel size (MambaATT)	3
Dropout rate (MambaATT)	0.1
Number of MambaATT layers	2

RESULTS AND DISCUSSION

The results demonstrate a notable performance improvement, with the CNN_ViT_MambaATT model achieving an accuracy of 0.92. This signifies improvement over

earlier deep learning research Bechelli and Delhommelle (2022) using the same ISIC dataset, which reported an accuracy of 0.87. The hybrid model improves skin cancer classification accuracy, aligning with new findings that highlight the value of intelligent fusion frameworks in skin lesion analysis (Dorathi Jayaseeli et al., 2025). Pon Selchiya et al. (2023) and Waheed et al. (2023) two models are proposed with 90% and 91% accuracy, respectively. The state-of-the-art (SOTA) model's benchmarks have been compared with our proposed model to show the accuracy of CNN-Vit_MambaATT as displayed in Table 3.

A confusion matrix is included to visualise classification outcomes. The model correctly identifies 589 benign and 603 malignant samples, with minor confusion between visually similar cases (Figure 8). CNN_ViT_MambaATT achieves competitive performance 92% while maintaining higher efficiency and lower parameter complexity. Misclassified cases mainly resulted from image blur or artifacts. Future work will incorporate hair removal and colour normalisation to improve results further. To ensure a comprehensive evaluation, additional metrics were introduced: specificity and a combined receiver operating characteristic (ROC) with area under the ROC curve show that CNN_ViT_MambaATT achieves an AUC of 0.945, as illustrated in Figure 9, which confirms its superior discrimination.

Table 3
Comparison proposed with SOTA methods

Model	Accuracy	Precision	Recall	F1-Score
(Huang et al., 2024)	0.8780	0.8760	0.8790	0.8770
(Naz et al., 2025)	0.8652	0.7645	0.9069	0.8293
(Lea, 2025)	0.9125	**0.9127**	0.9125	0.9125
Proposed (CNN_ViT_MambaATT)	**0.9200**	0.9100	**0.9150**	**0.9180**

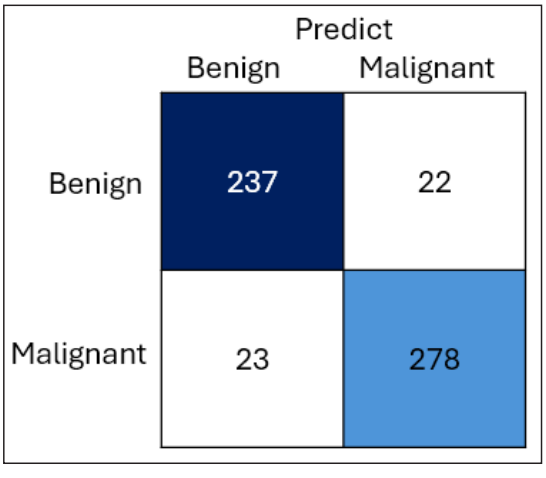

Figure 8. Confused matrix for testing the benign and malignant of dataset

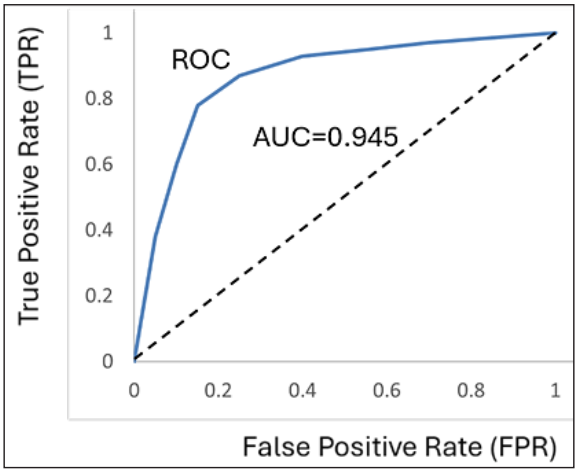

Figure 9. Area under the ROC curve of the proposed model

Ablation Experiment

The ablation experiment is a technique employed to assess the influence of specific components or features on model performance by systematically eliminating them from the model. It is typically applied to investigate the rationality of model design. The results of the ablation experiment for the proposed model are shown in Table 4.

The model achieved an accuracy of 92%. Embedding the MambaATT model resulted in a 1.4% accuracy to show that it improved classification precision through CNN-based presentation of image which counteracted the lack shortcoming from previous work. With the ViT model included, accuracy improved by 2.4%, demonstrating that by introducing global contextual information into models, ViT could improve the accuracy of classification task. Data augmentation was shown to provide a 1.6% improvement on the model, where we were able to generate more training examples through rotation, flipping, cropping and

Table 4
Results of the ablation experiment for the proposed model

Model	Accuracy	Precision	Recall	F1-Score
ResNet50	0.8610	0.8600	0.8580	0.8500
EfficientNet	0.8400	0.8430	0.8430	0.8380
ViT	0.8750	0.8740	0.8740	0.8740
Resnet50 + EfficientNet	0.8660	0.8600	0.8620	0.8630
Resnet50 + EfficientNet + ViT	0.8900	0.8850	0.8890	0.8800
Resnet50 + EfficientNet + ViT + MambaATT	0.9040	0.9010	0.8980	0.8950
(Resnet50 + EfficientNet + ViT + MambaATT + Data Augmentation)	**0.9200**	**0.9140**	**0.9180**	**0.9150**

so forth. The additional training examples helped the model to better learn the distribution and properties of the data leading to an improved performance. The performance measures for the various model configurations are outlined in Figure 10. To justify the effectiveness of data augmentation techniques, we evaluated the pre- and post-augmented results of ResNet50, ViT, and CNN_ViT_MambaATT models. As can be seen from Table 5, data augmentation significantly improves the model performance. This further reinforces the idea that deep learning models typically require large datasets to be trained, and small datasets may not sufficiently utilise their potential.

The research questions defined in the Introduction were effectively addressed. The CNN_ViT_MambaATT model showed better results than pure CNN or ViT networks. The module MambaATT improved the discriminative power by discovering the useful contextual transitions. The proposed feature fusion rule strengthened the robustness

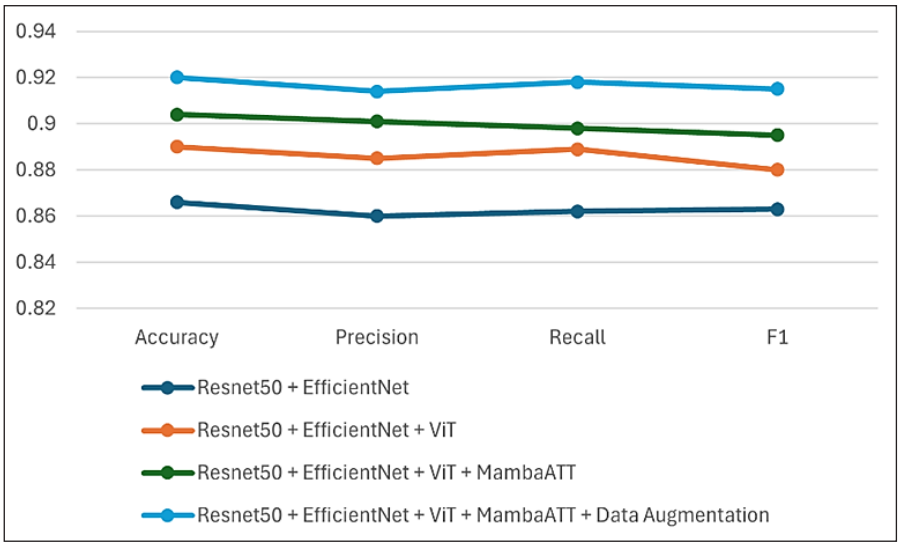

Figure 10. Performance metrics by model configurations

Table 5
Accuracy comparison of models before and after applying data augmentation

Model	Accuracy	Precision	Recall	F1-Score
Resnet50	0.8500	0.8450	0.8480	0.8490
(Resnet50+ Data Augmentation)	0.8610	0.8600	0.8640	0.8610
ViT	0.8600	0.8520	0.8560	0.8520
(ViT+ Data Augmentation)	0.8750	0.8680	0.8720	0.8680
CNN_ViT_MambaATT	0.9000	0.8910	0.8960	0.8920
(CNN_ViT_MambaATT +Data Augmentation)	**0.9200**	**0.9140**	**0.9180**	**0.9150**

of less data using the complementary spatial and global cues. This framework provides a basis for incorporating state-space models into medical image pipelines.

Misclassified samples were reviewed to search for common causes of the errors, such as low contrast of a lesion, overlapping artifacts, and visual similarities between early-stage melanoma and benign lesions. The Grad-CAM visualisation was used for interpreting the model's decision-making process, and it is evident that the proposed hybrid CNN–ViT–MambaATT architecture primarily emphasises lesion boundaries. In some cases, however, the activation of attention was also in harmless neighbouring regions, which might suggest over-sensitivity to background texture. Overall, the AUC of 0.945 and classification accuracy of 92% demonstrated a strong diagnostic performance of the model. Still, these perfusion-based measures should be used carefully in clinical practice. False negatives could result in missed melanomas, while false positives would warrant unnecessary surgical excisions. Therefore, this study frames the proposed model as a triage support tool that aids but does not substitute expert clinical judgment of dermatologists. In the future, we will work toward collaborating with dermatologists for validation and incorporating the model into clinical databases for evaluation of real-world reliability and safety.

CONCLUSION

In this study, we propose a novel model, CNN_ViT_MambaATT for skin cancers classification by utilising advanced DNN architectures such as ResNet50, VIT and MambaATT. This study also analysed the binary categorisation between benign and malignant dermoscopic skin cancer images. The skin cancer classification model CNN_VIT_MAMBAATT was trained and tested with image data from the ISIC dataset. The proposed method showed significant enhancements in malignant skin lesions and had 92% accuracy on the ISIC dataset. Conclusion: Classification focuses on how to couple learning mechanisms with the new attention mechanisms for improved classification accuracy and robustness. Furthermore, using the data-enlargement technique helped to overcome issues associated with small sample size and, in essence allowed the model to 'normalise' the data. We think that these results provide valuable understanding of the potential of hybrid models for medical image classification and encourage further developments towards reliable clinical systems.

LIMITATIONS AND FUTURE WORKS

Even though advances have been made, tasks such as denoising and efficiency are still necessary to balance the dermoscopic dataset. A potential solution incorporates a metaheuristic algorithm for deep renovation, optimal hyperparameter configuration advanced preprocessing methods to deal with these issues.

Hybrid models will particularly be a major direction in future studies for better explanation by visualisation tools, that is, Grade-Cam (Gamage et al., 2024), and for acknowledgements of the real world to make clinical relevance. VIT enriched with our architecture has strong prospects in skin cancer classification. In conjunction, hybrid methods to break pointing strategies and excellent preterm refined transition mechanisms that address current limitations, make the path towards more effective CAD systems, a clinically acceptable data control diagnostic (CAD) system. Future work should also focus on the slow processing speed of the Mamba Vit model. Though skin cancer detection can greatly benefit from high accuracy, its computation is slow and hence not within the reach of a wider population. Advancements in such models will have to entail architectural changes, simplified model structure, and faster training and estimation procedures. Furthermore, using the hardware acceleration of GPUs or TPUs increases performance.

The next study model is designed to be a compromise between computational efficiency and accuracy, as well as a practical, scalable clinical device. The objective will be to develop a hybrid system that combines the best attributes of Mamba and the VIT architecture, together with a model that is well-established for high-speed computations. In future work, we expect to validate the interpretability using clinically trained Dermatologists on Grad-CAM (Gamage et al., 2024) visualisations. train the proposed model on a large dataset, e.g., validate it on the ISIC 2019/2020 datasets. There will be some areas of future work integrating lesion-specific preprocessing hair removal, colour normalisation, testing on ISIC 2019 and ISIC 2020 benchmarks for wider generalisation and optimising computation cost for real clinical use.

ACKNOWLEDGEMENT

The authors wish to acknowledge the Faculty of Computer Science and Information Technology at UPM for the support and funding provided in this work.

REFERENCES

Abbas, Q., Daadaa, Y., Rashid, U., & Ibrahim, M. E. A. (2023). Assist-Dermo: A lightweight separable vision transformer model for multiclass skin lesion classification. *Diagnostics, 13*(15), Article 2531. https://doi.org/10.3390/diagnostics13152531

Ali, A., Shahbaz, H., & Damaševičius, R. (2025). xCViT: Improved vision transformer network with fusion of CNN and Xception for skin disease recognition with explainable AI. *Computers, Materials & Continua, 83*(1), 1367-1398. https://doi.org/10.32604/cmc.2025.059301

Ali, K., Shaikh, Z. A., Khan, A. A., & Laghari, A. A. (2022). Multiclass skin cancer classification using EfficientNets-a first step towards preventing skin cancer. *Neuroscience Informatics, 2*(4), Article 100034. https://doi.org/10.1016/j.neuri.2021.100034

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), Article 415. https://doi.org/10.3390/info14070415

Baykal Kablan, E., & Ayas, S. (2024). Skin lesion classification from dermoscopy images using ensemble learning of ConvNeXt models. *Signal, Image and Video Processing, 18*(8-9), 6353-6361. https://doi.org/10.1007/s11760-024-03321-y

Bechelli, S., & Delhommelle, J. (2022). Machine learning and deep learning algorithms for skin cancer classification from dermoscopic images. *Bioengineering, 9*(3), Article 97. https://doi.org/10.3390/bioengineering9030097

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2020). Generative pretraining from pixels. In H. Daumé III & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning* (Vol. 119, pp. 1691-1703). PMLR. https://proceedings.mlr.press/v119/chen20s.html

Dorathi Jayaseeli, J. D., Briskilal, J., Fancy, C., Vaitheeshwaran, V., Patibandla, R. S. M. L., Syed, K., & Swain, A. K. (2025). An intelligent framework for skin cancer detection and classification using fusion of squeeze-excitation-densenet with metaheuristic-driven ensemble deep learning models. *Scientific Reports, 15*(1), Article 92293. https://doi.org/10.1038/s41598-025-92293-1

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*. https://doi.org/10.48550/arXiv.2010.11929

Flosdorf, C., Engelker, J., Keller, I., & Mohr, N. (2024). Skin cancer detection utilising deep learning: Classification of skin lesion images using a vision transformer. *arXiv*. https://doi.org/10.48550/arXiv.2407.18554

Gamage, L., Isuranga, U., Meedeniya, D., De Silva, S., & Yogarajah, P. (2024). Melanoma skin cancer identification with explainability utilising mask guided technique. *Electronics, 13*(4), Article 680. https://doi.org/10.3390/electronics13040680

Goyal, M., Knackstedt, T., Yan, S., & Hassanpour, S. (2020). Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. *Computers in Biology and Medicine, 127*, Article 104065. https://doi.org/10.1016/j.compbiomed.2020.104065

Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modelling with selective state spaces. *arXiv*. https://doi.org/10.48550/arXiv.2312.00752

Han, S. S., Moon, I. J., Lim, W., Suh, I. S., Lee, S. Y., Na, J.-I., Kim, S. H., & Chang, S. E. (2020). Keratinocytic skin cancer detection on the face using region-based convolutional neural network. *JAMA Dermatology, 156*(1), 29-37. https://doi.org/10.1001/jamadermatol.2019.3807

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90

Huang, H., Hsu, B. W., Lee, C., & Tseng, V. S. (2021). Development of a lightweight deep learning model for cloud applications and remote diagnosis of skin cancers. *The Journal of Dermatology, 48*(3), 310-316. https://doi.org/10.1111/1346-8138.15683

Huang, S., Lei, H., Yang, J., Hang, T., Jin, L., Yao, Y., & Li, C. (2024). An attention mechanism and ensemble learning based on dermoscopic image classification. *Journal of Imaging Science and Technology, 68*(4), Article 40401. https://doi.org/10.2352/J.ImagingSci.Technol.2024.68.4.040403

Khan, M. A., Zhang, Y.-D., Sharif, M., & Akram, T. (2021). Pixels to classes: Intelligent learning framework for multiclass skin lesion localisation and classification. *Computers & Electrical Engineering, 90*, Article 106956. https://doi.org/10.1016/j.compeleceng.2020.106956

Kittler, H., Pehamberger, H., Wolff, K., & Binder, M. (2002). Diagnostic accuracy of dermoscopy. *The Lancet Oncology, 3*(3), 159-165. https://doi.org/10.1016/S1470-2045(02)00679-4

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097-1105.

Lea, S. (2025). Malignant vs. benign skin cancer classification using topological data analysis and machine learning. In *2025 IEEE 9th Forum on Research and Technologies for Society and Industry (RTSI)* (pp. 102-107). IEEE. https://doi.org/10.1109/RTSI64020.2025.11212537

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 255-258). MIT Press.

Meedeniya, D., De Silva, S., Gamage, L., & Isuranga, U. (2024). Skin cancer identification utilising deep learning: A survey. *IET Image Processing*. Advance online publication. https://doi.org/10.1049/ipr2.13219

Mikołajczyk, A., & Grochowski, M. (2018). Data augmentation for improving deep learning in image classification problem. In *2018 International Interdisciplinary PhD Workshop (IIPhDW)* (pp. 117-122). IEEE. https://doi.org/10.1109/IIPHDW.2018.8388338

Mobiny, A., Singh, A., & Van Nguyen, H. (2019). Risk-aware machine learning classifier for skin lesion diagnosis. *Journal of Clinical Medicine, 8*(8), Article 1241. https://doi.org/10.3390/jcm8081241

Naz, N. S., Mehmood, M. H., Ahmed, F., Ahmad, M., Rehman, A. U., Ismael, W. M., & Adnan, K. M. (2025). Privacy preserving skin cancer diagnosis through federated deep learning and explainable AI.

Nie, Y., Sommella, P., Carratù, M., O'Nils, M., & Lundgren, J. (2023). A deep CNN-transformer hybrid model for skin lesion classification of dermoscopic images using focal loss. *Diagnostics, 13*(1), Article 72. https://doi.org/10.3390/diagnostics13010072

Ozdemir, B., & Pacal, I. (2025). A robust deep learning framework for multiclass skin cancer classification. *Scientific Reports, 15*(1), Article 4938. https://doi.org/10.1038/s41598-025-89230-7

Pon Selchiya, R., Manimegalai, P., & Thomas George, S. (2023). Comparative analysis-based melanoma detection in dermoscopic images with deep learning techniques. In *2023 5th International Conference on Electrical, Computer and Communication Technologies (ICECCT)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICECCT56650.2023.10179820

Reis, H. C., & Turk, V. (2024). Fusion of transformer attention and CNN features for skin cancer detection. *Applied Soft Computing, 164*, Article 112013. https://doi.org/10.1016/j.asoc.2024.112013

Siegel, R. L., Kratzer, T. B., Giaquinto, A. N., Sung, H., & Jemal, A. (2025). Cancer statistics, 2025. *CA: A Cancer Journal for Clinicians, 75*(1), 10-45. https://doi.org/10.3322/caac.21871

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)* (pp. 6105-6114). PMLR.

Tay, Y., Dehghani, M., Gupta, J., Bahri, D., Aribandi, V., Qin, Z., & Metzler, D. (2021). Are pre-trained convolutions better than pre-trained transformers? *arXiv*. https://doi.org/10.48550/arXiv.2105.03322

Waheed, S. R., Saadi, S. M., Rahim, M. S. M., Suaib, N. M., Najjar, F. H., Adnan, M. M., & Salim, A. A. (2023). Melanoma skin cancer classification based on CNN deep learning algorithms. *Malaysian Journal of Fundamental and Applied Sciences, 19*(3), 299-305. https://doi.org/10.11113/mjfas.v19n3.2900

Yang, G., Luo, S., & Greer, P. (2025). Boosting skin cancer classification: A multi-scale attention and ensemble approach with vision transformers. *Sensors, 25*(8), Article 2479. https://doi.org/10.3390/s25082479

Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., & Wang, X. (2024). Vision Mamba: Efficient visual representation learning with bidirectional state space model. *Proceedings of Machine Learning Research, 235*, 62429-62442. https://proceedings.mlr.press/v235/zhu24f.html.