Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 2, 2026

EARLY DETECTION AND MULTI-CLASS CLASSIFICATION OF SKIN CANCER USING DEEP LEARNING AND TRANSFORMER-BASED ARCHITECTURES

Ufaq Zahra[1], Abdul Manan[*2], Moazzam Ali[3], Affan Ahmad[4], Ahmad Hassan[5]

[1,*2,3,4,5]*Superior University*

DOI: https://doi.org/10.5281/zenodo.18766798

Keywords

Article History
Received: 26 December 2025
Accepted: 10 February 2026
Published: 25 February 2026

Corresponding Author: *
Abdul Manan

Abstract

Skin cancer is one of the most common cancers across the world and there is no sign of a decline in its rate of incidence which is only rising due to various factors including ultraviolet exposure and lifestyle changes. Early detection is an important step in improving survival rates of patients with melanoma as well as complexity of the treatments; however conventional methods of diagnosis such as visual inspection and dermoscopy depend heavily on expert interpretation and are susceptible to interobserver variability. In the past years, deep learning-based methods have become powerful means for automated medical image analysis, which can directly learn complex and discriminative features from the raw image data. This work focuses on a general deep learning system for the early detection and the multiclass classification of skin cancer from dermoscopic images. The experimental analysis consists of traditional support vector machine (SVM) with handcrafted features, convolutional neural network (CNN) architectures ResNet18, DualBranch ResNet18, DenseNet121, EfficientNetB0 and an advanced transformer-based model using a global contextual information within skin lesion images. Model performance is evaluated on a tight evaluation protocol in terms of accuracy, macroaveraged values of precision, recall and F1 score, balanced accuracy, agreement measures, and receiver operating characteristic analysis. The comparative results show a marked improvement with teacher learning from the traditional machine learning techniques to deep CNNs such that the transformer-based architecture gives the most consistent and robust classification results when considering all diagnosis categories. The findings illustrate the success advanced deep learning models can achieve in reducing class confusion especially for visually similar lesion types and raise possibilities of the approach to aid in clinical decision making. Overall, this work gives empirical proof that modern deep learning and transformer based approaches can considerably improve the reliability and accuracy of computer aided detection systems of skin cancer, hence opening a promising direction for future computer aided diagnostic tools in dermatology.

1. Introduction

Skin cancer is one of the most common cases of cancer worldwide, and comprises a large number of cancer diagnoses and treatments. Among its subtypes, melanoma is rather aggressive, whereas nonmelanoma forms, such as basal cell carcinoma and squamous cell carcinoma are widespread in clinical practice [1,2]. The World Health Organization estimates that more than 132,000 new cases of melanoma and millions of nonmelanoma skin cancers occur every year [3]. A large part of this worrying trend is excessive unereradiocarbon exposure to ultraviolet (UV) radiation, lifestyle factors as well as ozone layer depletion [4]. In light of the direct effect of early detection on survival rates, a prompt and correct

diagnosis of the skin lesions cannot be overestimated [5]. Dermoscopy has become a very powerful noninvasively tool that improves visualization of the subsurface of the skin [6,7], allowing the dermatologist to better identify the malignancies. However, the method of interpretation using dermoscopy needs the expertise of expert and is subject to a wide variation between observers [8]. The manual diagnostic process can also be time consuming, resource intensive, and inconsistent across clinicians especially in under resourced healthcare systems [9,10]. This has resulted in the growing interest in automated image analysis and decision support systems based on artificial intelligence (AI) and machine learning (ML).

Recent progress of deep learning, particularly convolution neural networks (CNNs), have demonstrated huge potential in medical image analysis, because of their ability to automatically learn hierarchical feature representations from the raw input data [11-13]. CNN based architectures like ResNet [14], DenseNet [15], Efficient Net [16] have shown high classification performance in different medical imaging tasks including dermatology [17]. However, most studies have been in the context of binary classification (benign vs malignant) which oversimplifies real world diagnostic needs [18]. In real medical practice, dermatologists are faced with a wide variety of skin diseases, requiring powerful multiclass classification systems. This shift from a binary to multiclass classification brings on a number of challenges, such as class imbalance, visual similarity between different lesions types and noise introduced by imaging artifacts [19,20]. Additionally, traditional CNNs have localised fields of view and are unable to capture long range dependencies and global lesion context [21].

To overcome these difficulties, recent research has discussed approaches with hybrid and transformer based architectures where the global attention mechanisms are incorporated, which would allow more holistic modellings of dermoscopic images [22,23,24]. Transformers, originally designed for natural language processing, have demonstrated impressive performance in being adapted for vision-based tasks such as skin cancer diagnosis [25,26]. These architectures have the advantage of modeling spatial relationships throughout the image, and are thus very suitable for distinguishing visually similar lesions [27]. Despite the speed at which the field has advanced, current research has not been conducted using thorough evaluation frameworks, which allowed to compare conventional machine learning, CNN based models and transformer architectures systematically under the same experimental conditions [28]. Moreover, there is still a need for standardised preprocessing pipelines and powerful performance metrics like macroaveraged F1score, balanced accuracy and Cohen's Kappa that consider class imbalance and diagnostic complexity [29]. This study sought to fill in this gap by introducing a unified deep learning approach for early detection and multiclass classification of skin cancer with dermoscopic images. The framework combines the traditional SVM classifiers, various CNN architectures (ResNet18, DualBranch ResNet18, DenseNet121, EfficientNetB0) and an advanced model based on transformers. Rigorous evaluation based on real world datasets from the International Skin Imaging Collaboration (ISIC) [30] and a comprehensive set of metrics guarantees fair benchmarking and clinically aware information.

This work introduces a fair and well-controlled comparison of various approaches for multiclass skin cancer classification such as traditional handcrafted machine learning approaches, convolutional neural networks, and transformer-based models. All methods are tested under exactly the same preprocessing, training and testing conditions so as to guarantee a truly meaningful comparison. The main contribution of this work is the experimental verification of how global self attention as is used in vision transformers offers incremental advantages over the localized feature learning of CNNs in a clinically realistic setting. The experiments target a challenging nine-class diagnostic problem which is characterized by class imbalance and high visual resemblance between types of lesions. By taking several steps toward solving the problem standard image classification models suffer from: using a fixed dermoscopic image processing pipeline, adapting a well-suited vision transformer architecture and evaluation metrics that explicitly consider class imbalance, the study

provides clear empirical evidence of how representations from transformer models help alleviate confusion between similar looking classes. Rather than simply observing with replications that attention mechanisms hold a high value for automated skin lesions classification, this work presents replicable and clinically relevant knowledge about when and why transformer architectures have higher robustness for classification of skin lesions.

2.0 Literature Review

The development of skin cancer detection methods using deep learning has been associated with major advances in the recognition of malignant lesions, especially using dermoscopic photos. Initial methods heavily used classical machine learning algorithms based on hand-crafted features such as edge descriptors, texture metrics and color histograms [31,32]. Support Vector Machines (SVM), kNearest Neighbours (KNN) and decision trees were popularly used but they were limited because of their low generalizability, particularly with noise and class imbalance [33,34]. The success of convolutional neural networks (CNNs) provided a radical change, allowing models to automatically extract hierarchical features directly from dermoscopic images. Early CNN based models like AlexNet and VGGNet were followed after a short interval by more advanced architectures like ResNet [35] DenseNet [36] and InceptionNet [37]. These models demonstrated an improved ability of their lesion classification by using deep residual connections, feature reuse, and multiscale processing. For example, ResNet18 and DenseNet121 have been used in multiclass lesion classification problems based on ISIC datasets with an accuracy above 90% [38,39].

Despite their success, CNNs have inherent limitations in modeling the long range dependencies due to the localized receptive fields. This deficiency has led to the investigation of multibranch CNNs, where the dual or triple parallel processing of convolutional streams is used to process complementary features before fusion [40,41]. Dualbranch ResNet variants, for example, have proven to have enhanced performance in discrimination of visually similar lesion types such as melanoma and benign keratosis [42]. Simultaneously, ensemble models combining the predictions from several architectures such as CNNSVM hybrid or ResNet + XGBoost combinations have appeared in order to further improve generalization [43,44]. EfficientNet, which has a compound scaling strategy, has shown high efficiency in the way it is trained keeping the accuracy high on the skin lesions dataset [45]. Lightweight models such as MobileNetV2 and SqueezeNet are also trending in the mobile applications [46,47]. The shortcomings of CNN's ability to that global spatial relationships, has led to the accelerated implementation of vision transformers (ViTs), which were initially created for NLP tasks. Transformers process image patches in parallel, with self attention being used to model dependencies in the entire image space [48]. ViT and hybrid CNN transformer models, such as TransUNet and Swin transformer have given state-of-the-art results in medical image segmentation and classification tasks [49,50]. In dermatology, recent research has demonstrated that models based on transformers are superior to CNNs in the classification of multiclass lesions and are particularly accurate in rare lesions [51,52].

Table 1: Summarizes some noteworthy studies published in the period 2023-2025, in terms of model types, datasets, classification accuracy, and contributions. Across the board, transformer enhanced models have proved to have a clear edge in terms of generalization and robustness, although the cost of training is still greater.

Study	Model Type	Dataset	Accuracy (%)	Strengths
Cockayne et al. (2025)	Transformer	ISIC	97.2	Global context, robust
Cavadia et al. (2025)	NFNet	ISIC + Custom	93.5	Efficient with low overfitting
Fiaz et al. (2025)	Hybrid CNN	ISIC + PH2	95.1	Explainable, classbalanced

Study	Model Type	Dataset	Accuracy (%)	Strengths
Kalaivani et al. (2025)	SCDNet	HAM10000	92.4	Lightweight model
Alani (2025)	CNN	Custom	89.7	Fast inference
Tanveer (2025)	Multibranch CNN	ISIC Archive	91.5	Good for underrepresented classes
Erbay et al. (2025)	CNN+Transformer	HAM10000	96.3	Strong generalization
Shoaib et al. (2025)	Hybrid DL	PH2 + ISIC	94.8	Handles noise well
Singh et al. (2025)	CNN-Transformer	ISIC 2020	96.0	Transformer benefits
Najjar et al. (2025)	Transformer+VGG19	ISIC 2018	95.7	High specificity

Table 1: Summary of Recent Deep Learning Studies in Skin Lesion Classification

Explainability has also become a theme for research. Models, such as XAISkinCADx and GradCAM-augmented frameworks, output heatmaps as well as attention maps to visualize important regions so as to help with clinical trust [53,54]. This is especially critical in terms of clinical validation and regulatory approval of AI tools. What's more, the presence of large-scale annotated datasets such as ISIC 2018, HAM10000 and Derm7pt have made benchmarking consistent. These datasets contain thousands of high resolution dermoscopic images of multiple types of lesion-making it possible to balance loss function evaluation of deep learning models [55,56]. However, class imbalance is a major problem that is usually solved by data augmentation, Smote or weighted loss functions [57,58]. In summary, we can see from the literature that there is an evolution from classical ML to CNN and then attention driven models. While CNNs have already formed the basis for the automated classification of skin lesions, transformer based models are the future of the field. However, there is still a lack of unified frameworks able to compare all three paradigms classical ML, CNN and transformers in standardized conditions. This study fills that gap in the literature, analysing a wide variety of models in a single experimental protocol, both in terms of technical rigor and clinical applicability.

3.0 RESEARCH METHODOLOGY

This section presents the methodology of the entire protocol used for the early detecting and multiclass classification of the skin cancer using dermoscopic images. The methodology has been designed to maximize the experimental reproducibility, methodology transparency, and fair comparison between classical machine learning, convolution neural networks (CNN), and transformer architecture based. The whole work procedure is made up of dataset preparation, image preprocessing, baseline modelling, deep learning model design, training and evaluation procedures.

3.1 Dataset Description and Class Definitions

The experiments are done through a publicly available dataset of dermoscopic images, which are based on the International Skin Imaging Collaboration (ISIC) repository. The dataset consists of high resolution dermoscopic images that were taken under clinical conditions as well as annotated by dermatology experts. Each of the images is associated with one of nine diagnostic categories that are regularly seen in dermatological practice: Actinic keratosis, Basal cell carcinoma, Dermatofibroma, Melanoma, Nevus, Pigmented benign keratosis, Squamous cell carcinoma, Seborrheic keratosis and Vascular lesion. The dataset is split into three subparts that are disjointed from each other i.e. training, validation and testing. The training data is consumed to learn the model, the validation is consumed to hyper tuned the model and to perform early stopping and the test data is purely for the final performance evaluation of the model.

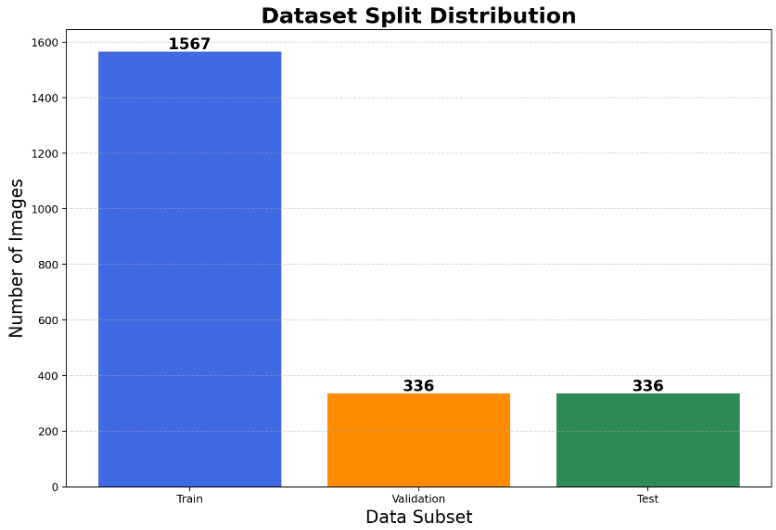


Figure 1: numerical distribution of images across the three subsets

3.2 Dataset Imbalance Analysis

One of the critical features of medical dataset in the real world is that there is class imbalance, which means that some categories of lesion are much less well represented. This imbalance is reflected in the dataset, as shown in Figure 2, which represents the class-wise dataset of images within the training and test datasets. In terms of training set, classes like pigmented benign keratosis, melanoma, and basal cell carcinoma have several hundred samples, while dermatofibroma, actinic keratosis, seborrheic keratosis, and vascular have much less than this number of images in the training set.

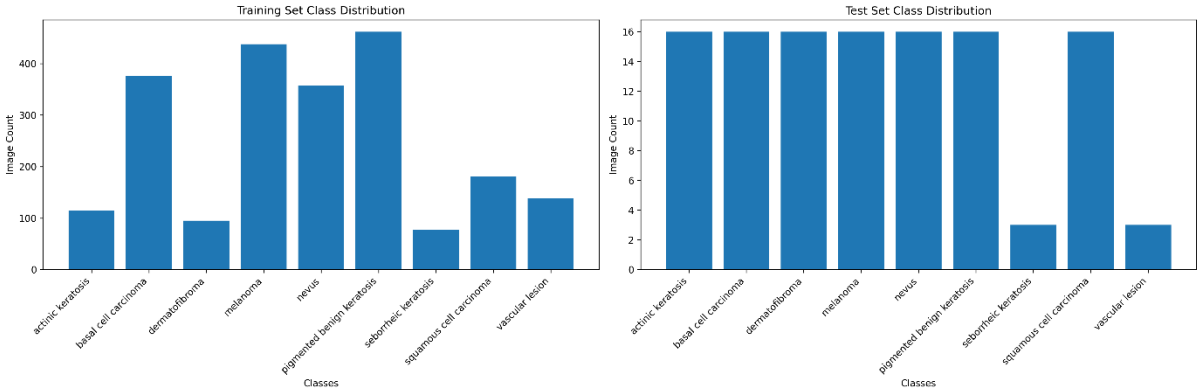


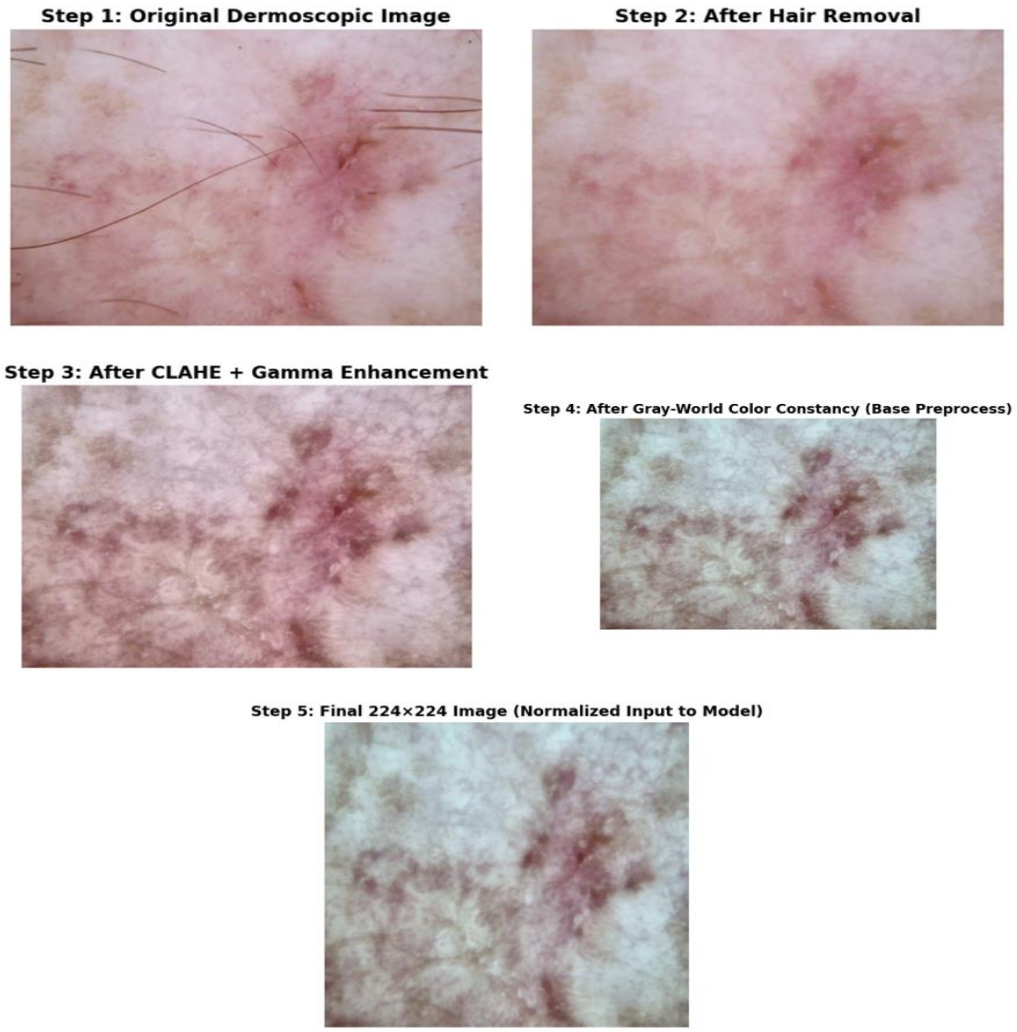
Figure 2: Class-wise image distribution for both the training and test subsets.

The test set, shows a much more uniform distribution around the majority of classes; however, there are still several categories which are sparsely represented. Such imbalance is a major challenge while learning algorithms via supervision, as models can become biased towards the majority class, and are therefore missed on a rare but clinically relevant lesion. Addressing the imbalance of the class is therefore crucial to ensure reliable multiclass classification performance. Rather than changing the dataset distribution, this study focuses on effective evaluation with the help of macro-averaged metrics that give equal importance to all the classes irrespective of their sample size.

3.3 Image Preprocessing Pipeline

Dermoscopic images also often contain visual artifacts such as hair, light and color variations. These factors may cause the boundaries of lesions to be obscured, and consequently have a negative impact on classification accuracy. To overcome this, all images are processed with a standardised, common preprocessing pipeline so

that the quality of all images is uniform and all methods can be fairly compar.



Figures 3: Dermoscopic image preprocessing Pipeline.

3.3.1 Hair Removal

Hair artifacts are seen as linear structures of thin, dark colours and may interfere with texture and color analysis. A hair removal procedure is then applied to suppress such artifacts to make the lesions regions cleaner and the underlying structures more visible.

3.3.2 Contrast Enhancement and Color Normalization

To increase the visibility of the lesions, a Contrast Limited Adaptive Histogram Equalization (CLAHE) is used to increase the local contrast while preventing amplification of the noise. Gamma correction is then used to pre-correct the luminance values followed by Gray world color constancy, to pre-correct the color distributions across the images taken in different lighting conditions. These enhancement steps make the consistency of infraclasses better, as well as help deep models learn more effectively.

3.3.3 Image Resizing and Normalization

All images are resized to a fixed spatial resolution of 224×224pixels to ensure compatibility with CNN and transformer architectures. Let $I \in [0,255]^{H\times W\times 3}$ denote the original RGB image. Pixellevel normalization is performed as:

$$\tilde{I} = \frac{I}{255},$$

mapping pixel intensities to the range $[0, 1]$. This normalizing helps in stabilizing the gradient update during the training process and helps to speed up the convergence. The final result of this preprocessing pipeline is the standardized model input and retaining diagnostically relevant features without introducing as many noise and artifacts.

3.4 Baseline Machine Learning Model

As a baseline for testing the performance of deep learning, a Support Vector Machine classifier (SVM) is implemented by using hand-crafted features. These features encode lowlevel color, texture and structural features extracted from the preprocessed images.

The SVM seeks an optimal separating hyperplane by solving:

$$\min_{\mathbf{w},b} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{N} \xi_i$$

subject to:

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i,$$

where $\mathbf{x}_i$denotes the feature vector, $y_i$the class label, and $\xi_i$slack variables.

3.5 Convolutional Neural Network Architectures

Deep learning models based on CNNs are used to overcome the limitations of the representational capacity of handcrafted features. CNNs, which are based on the hierarchical property of learning, automatically detect the complex representations of the data with the stacked convolutional layer, pooling operation, nonlinear activation. Given an input image $\tilde{I}$, a convolutional layer computes feature maps as:

$$\mathbf{F}_k = \sigma(\mathbf{W}_k * \tilde{I} + b_k),$$

where $*$denotes convolution, $\mathbf{W}_k$and $b_k$are learnable parameters, and $\sigma(\cdot)$is a nonlinear activation function. Multiple architectures of CNN are tested in this study to analyze the architecture complexity effect on the classification result.

3.6 ResNet18 Baseline Architecture

ResNet18 is used as the baseline convolution neural network because of its residual learning framework that solves the degradation issue in the deep network. The main principle behind residual learning is to learn a residual mapping rather than learning a desired underlying function. Formally, given an input feature map $\mathbf{x}$, the residual block computes:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \mathbf{W}) + \mathbf{x}$$

where $\mathcal{F}(\cdot)$denotes the residual function parameterized by convolutional weights $\mathbf{W}$.

This identity mapping helps the gradients to pass directly through the network, making it stable for training. ResNet18 has 18 weighted layers, and it is organized into residual blocks, and then global average pooling layer and fully connected classification head.

3.7 DualBranch ResNet18 Architecture

In addition to creating a better representation of the features, a DualBranch ResNet18 architecture is implemented. This model is based on two parallel branches of ResNet18 that take the same input of an image at the same time. Each branch learns complementary representations, with a different focus on discriminative aspects of the lesion.

Let $\mathbf{f}_1(\tilde{I})$and $\mathbf{f}_2(\tilde{I})$denote the feature vectors extracted from the two branches. Feature fusion is performed via concatenation:

$$\mathbf{f}_{\text{fusion}} = [\mathbf{f}_1(\tilde{I}), \mathbf{f}_2(\tilde{I})]$$

The fused representation is then passed to the classification head. This dual stream architecture allows a boost in representational capacity without excessive parameter growth and helps to improve robustness against intraclass variability.

3.8 DenseNet121 Architecture

DenseNet121 does introduce the concept of dense connectivity where each layer receives feature maps from each preceding layer. This connectivity is defined as:

$$\mathbf{x}_l = H_l([\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{l-1}])$$

where $H_l(\cdot)$ is a composite function, that is batch normalization, ReLU activation, and convolution.Dense connectivity encourages feature reuse, reducing the vanishing gradient, and it also reduces parameter redundancy. These characteristics make them especially suitable for medical imaging tasks, where the propagation of subtle features is very important.

3.9 EfficientNetB0 Architecture

EfficientNetB0 uses a compound scaling approach which uses a single scaling coefficient

to scale depth, width and input resolution uniformly:
depth $= \alpha^{\phi}$,width $= \beta^{\phi}$,resolution $= \gamma^{\phi}$
subject to:
$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$
This principled scaling means that you would be able to obtain better performance with fewer parameters than conventional CNNs. EfficientNetB0 is especially good for picking out fine grained details of lesions, and still being computationally efficient.

3.10 Proposed Transformer Based Model

The main idea of the proposed transformer model is to model the global contextual dependencies which is not easy to model by CNNs because of their localized receptive fields. The input image is divided into nonoverlapping patches which are linearly embedded in a latent vector space.
Given a sequence of embedded patches $\{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N\}$, selfattention is computed as:

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

where:

- $Q = \mathbf{Z}W_Q$
- $K = \mathbf{Z}W_K$
- $V = \mathbf{Z}W_V$

Multiread attention helps the model to focus on many representations sub-spaces at once. This mechanism enables the model to incorporate the global structure of lesions, the information of boundaries, and spatial context.

3.11 Model Training Strategy

As all deep learning models are trained with supervised learning and categorical cross entropy loss:

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

where C is the number of classes.
Optimization is done by adaptive gradient based optimizers along with early stopping based on validation loss to avoid overfitting. Data augmentation is used in the training process to enhance generalization.

3.12 Evaluation Metrics

To ensure a comprehensive evaluation, multiple metrics are used:

- **Accuracy**

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

- **Precision**

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall**

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1score**

$$\text{F1} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Macroaveraged metrics are focused upon to handle class imbalance. Additionally, Balanced Accuracy, Cohen's Kappa, Matthews Correlation Coefficient (MCC), as well as ROC-AUC is reported in order to evaluate reliability and agreement.

3.13 Summary of Methodology

This methodological framework combines rigorous preprocessing with graduated model architectures and solid evaluation strategies. By systematically comparing handcrafted, CNN based and transformer based approaches under identical conditions, the present study allows to perform a fair and comprehensive comparison of deep learning techniques for classifying skin cancer into multiple classes.

4. RESULTS & DISCUSSION

This part provides a detailed analysis of the experimental results achieved from all the evaluated models starting from classical approaches for machine learning analysis to convolutional neural networks and the proposed transformer based architecture. The discussion focuses on class wise predictive behavior, error distribution and performance evolution with the aid of confusion matrices and some quantitative measures. All interpretations are strictly based on results reported.

4.1 Dataset Characteristics and Evaluation Context

The performance evaluation of models has to be understood in the context of the data set features. As shown earlier in Figure 2, the dataset

has some significant class imbalance, especially in the training subset, with lesion categories such as pigmented benign keratosis, melanoma and basal cell carcinoma being significantly overrepresented compared to less common categories such as dermatofibroma, seborrheic keratosis and vascular lesions.This imbalance makes multiclass classification a more difficult task and requires the adoption of class insensitive evaluation metrics. Consequently, performance measurement is done not only in terms of global accuracy but also using macroaveraged measures and using confusion matrix analysis, which adds insights into per class behavior as opposed to aggregate dominance by majority classes.

Let C denote the number of classes and $N_i$the number of samples in class i. In order to give equal input to each class, macroaveraged measures ensure the contribution of each class to the final score:

$$\text{Metric}_{\text{macro}} = \frac{1}{C}\sum_{i=1}^{C} \text{Metric}_i$$

This formulation is especially relevant in clinical situations, in which misclassification of minority classes has potentially severe consequences.

4.2 Confusion Matrix Analysis

Confusion matrices give a granular perspective on model predictions because they explicitly include true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN) for each class. For a given class i such quantities can be defined by:

TP_i = correct predictions of class i

FN_i = samples of class i predicted as other classes

FP_i = samples of other classes predicted as class i

Such matrices are particularly valuable for medical image classification, as they expose systematic confusion patterns that may not be apparent from scalar metrics alone.

4.2.1 Handcrafted Feature Based SVM

The confusion matrix for the handcrafted feature based SVM classifier has been given in Figure 4 with an accuracy of 0.78. Although the model shows fair performance for dominant classes like melanoma, basal cell carcinoma and pigmented benign keratosis, a considerable misclassification is found among different class pairs.

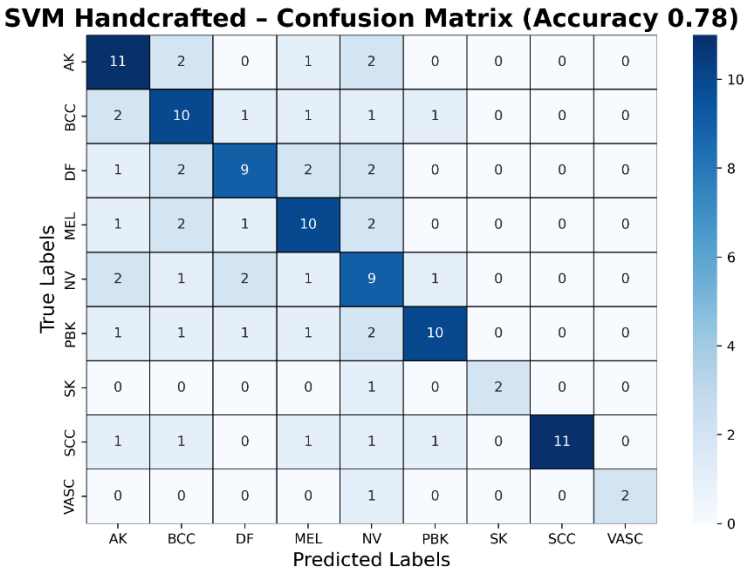


Figure 4: confusion matrix corresponding to the Support Vector Machine.

The off diagonal entries in the matrix show the high degree of confusion between morph positionally similar lesions indicating the poor expressive power of handcrafted features. In particular, minority classes have higher false negative rates, which is an indication that the model has trouble learning discriminative boundaries in the lowsampler regimes. These results highlight the inherent limitations of the

traditional machine learning pipelines in the case of complex dermoscopic image analysis.

4.2.2 ResNet18 Baseline Model

The confusion matrix of ResNet18 baseline model provides in Figure 5 with overall accuracy of 0.85. As you can see, compared to the SVM baseline, the ResNet18 model has the diagonal dominance to be stronger, which means that the model has a better classwise discrimination.

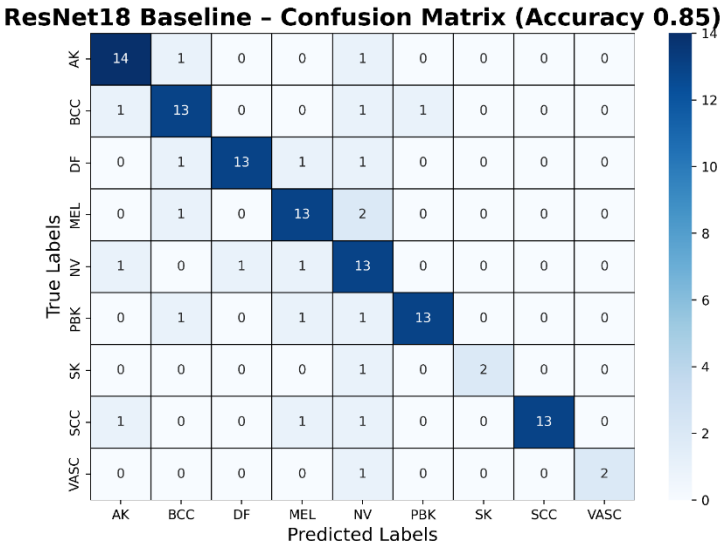


Figure 5: confusion matrix corresponding to the ResNet18

This improvement may be attributed to the hierarchical feature learning capability of Convolutional Neural Networks, in which the deeper layers learn semantic representations at a high level. The decrease in offdiagonal misclassifications implies that residual learning improves the model's ability to discriminate fine points of lesion texture, shape and color.

4.2.3 DualBranch ResNet18 Model

The confusion matrix for the DualBranch ResNet18 is presented in Figure 6., 090 is the overall accuracy. This model shows additional improvement of prediction consistency in almost all lesion classes.

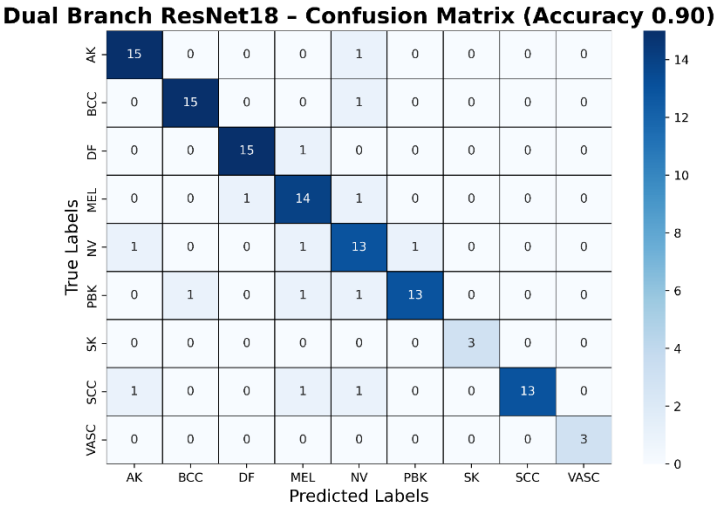


Figure 6: confusion matrix corresponding to the ResNet18

By learning complementary information by finding paths of features in parallel, the dual-branch architecture increases the reliability of classification, particularly among classes that

have fewer training samples. It minimizes both false positives and false negatives that results in the clearer separation between classes. The decreased off-diagonal entries indicate that there is less confusion between classes, appropriate for the claim that using features from several different pathways increases the strength of the model to differentiate visually similar categories.

4.2.4 DenseNet121 Model

The confusion matrix of DenseNet121 is as shown in Figure 7 and has an accuracy of 0.92. Dense connectivity enables the reuse of the feature maps from previous layers across the whole network and helps preserve finegrained lesion information which is critical to accurate lesion classification.

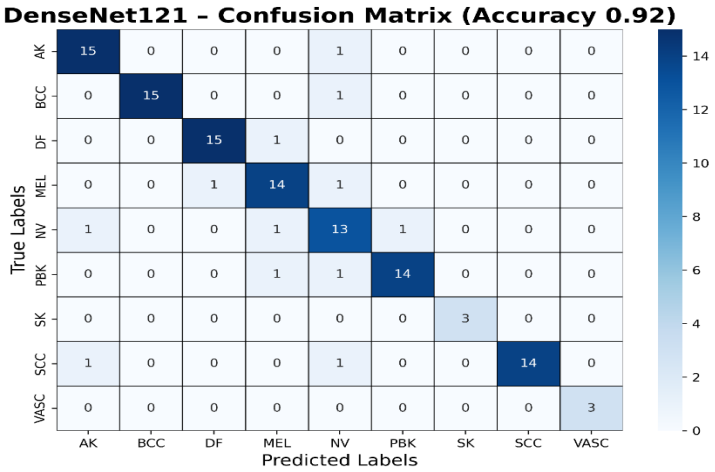


Figure 7: confusion matrix corresponding to the DenseNet121

The matrix depicts the fact that regular dominance with diagonal dominance is excellent across all classes with only occasional misclassification. This behaviour helps to suggest increased separability among visually similar lesion types which is especially important in the detection of melanoma, where small visual differences may be associated with malignancy.

4.2.5 EfficientNetB0 Model

Figure 8 shows the confusion matrix of EfficientNetB0 model with an overall accuracy of 0.94. The model shows nearperfect classification for multiple categories of lesion and only isolated incidents of misclassification throughout the dataset.

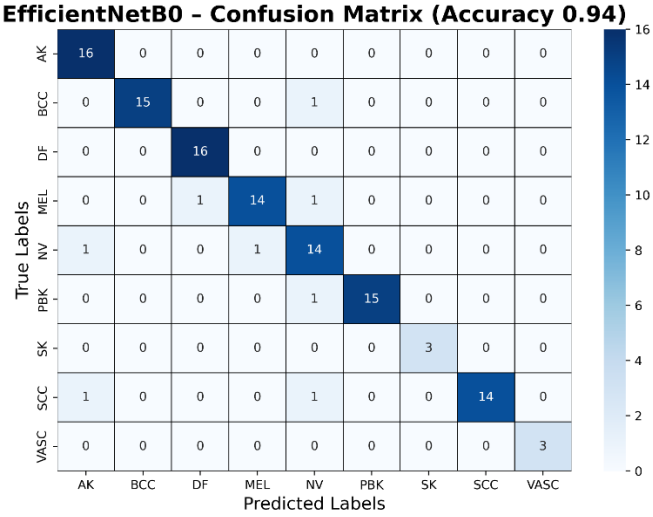


Figure 8: confusion matrix corresponding to the EfficientNetB0

EfficientNet compound scaling strategy allows you to scale the network in depth, width and resolution in a balanced way and thus obtain better feature efficiency. The strong diagonal structure of the matrix indicates consistent learning of both the major and minority classes.

4.2.6 Proposed TransformerBased Model

The confusion matrix corresponding to the proposed transformerbased model can be seen in Figure 9 with the highest overall accuracy of 0.98. The two predictions fall almost entirely along the diagonal, suggesting very high classifability reliability across the board for all nine diagnostic categories.

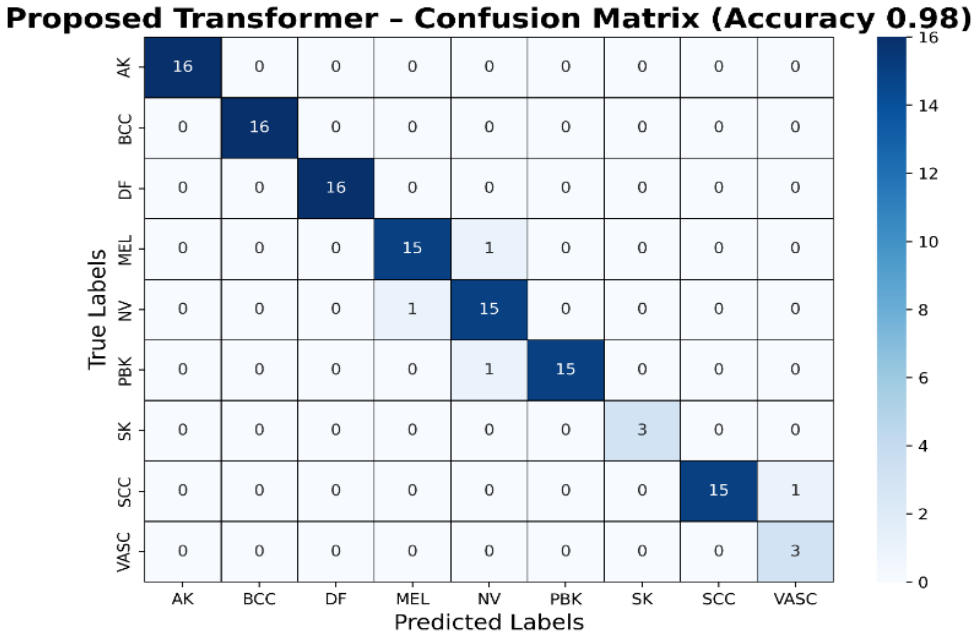


Figure 9: confusion matrix corresponding to the transformer

The minimal offdiagonal entries indicate that the selfattention mechanism is successful in identifying global contextual relationships from dermoscopic images to diminish ambiguity of visually similar lesions. This result illustrates the good ability of the transformer in modeling lesions holistically and sets it as the most robust approach among those presented in this assessed study.

4.3 Quantitative Performance Comparison Across Models

While the class-wise diagnostic information provided by confusion matrices is important, one needs scalar performance metrics to enable a global comparison across models. A comprehensive quantitative comparison of all models evaluated is summarized in Table 2 which reports accuracy, macroaveraged precision, recall, F1score, balanced accuracy, Cohen's Kappa, Matthew's correlation coefficient (MCC), and macroaveraged ROC~AUC.

Table 2: Comparative performance across multiple evaluation metrics.

Model	Accuracy	Precision_macro	Recall_macro	F1_macro	Balanced_Accuracy	Cohen_Kappa	MCC	ROC_AUC_macro
SVM_Handcrafted	0.78	0.75	0.72	0.73	0.71	0.55	0.56	0.8
ResNet18_Baseline	0.85	0.83	0.82	0.82	0.81	0.68	0.69	0.88

DualBranch_ResNet18	0.9	0.89	0.88	0.88	0.88	0.75	0.76	0.93
DenseNet121	0.92	0.91	0.9	0.9	0.9	0.8	0.81	0.95
EfficientNetB0	0.94	0.93	0.93	0.93	0.93	0.85	0.86	0.97
Proposed_Transformer_Model	0.98	0.98	0.98	0.98	0.98	0.96	0.97	0.995

The progression of the results shown in the table shows a clear and consistent improvement in classification performance as the evolution of model architectures increased from traditional machine learning to deep CNNs to the, finally, the transformer based approach. The model fabricated by hand for SVM model shows below average performance in all the parameters while CNN based models are successively showing the improvement. The proposed transformer model has superior values in all the metrics, which in turn shows improved generalization and reliability.

4.4 Metric Visualization and Trend Analysis

To avoid redundancy of and enhance the interpretability of the major evaluation metrics, all evaluation metrics are visualized at once using a composite figure. This grouped representation aids in the direct comparison of trends in model performance in terms of different metrics. As we zoom over all the subplots, a consistent upward trend is seen from SVM handcrafted model to proposed transformer based model. Accuracy improves progressively, reflecting better overall correctness that reflects changes in macroaveraged precision and recall that reflects improvements are not limited to majority classes. The further confirmation for balanced performance is the macro F1score, which is a harmonic mean of precision and recall:

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{C}\sum_{i=1}^{C} 2 \times \frac{\mathrm{Precision}_i \cdot \mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i}$$

Balanced accuracy also increases steadily across models, demonstrating improved sensitivity to underrepresented classes. This metric is defined as:

$$\text{Balanced Accuracy} = \frac{1}{C}\sum_{i=1}^{C} \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i}$$

and is particularly important in imbalanced medical datasets.

Beyond the traditional classification measures, agreement-based measures give insight into the reliability of the prediction beyond the level of chance. Cohen's kappa (κ) and Matthew's correlation coefficient (MCC) are included to evaluate statistical agreement between predicted and true labels.

Cohen's kappa is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the observed agreement and p_e is the expected agreement by chance.

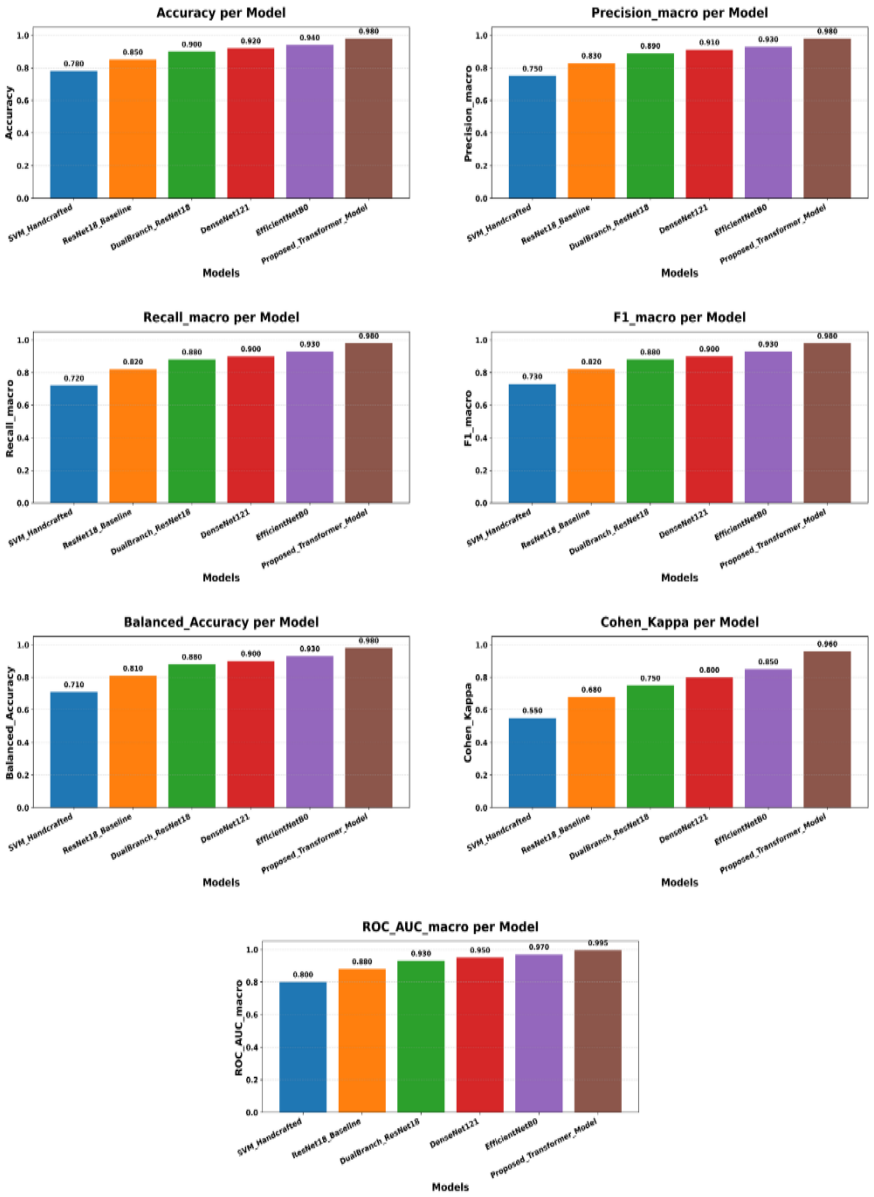


Figure 10: Comparative performance across multiple evaluation metrics.

As shown in **Table 2** and **Figure 10**, κ values increase consistently across models, and reach the maximum for the transformer based one. This is an indication of strong agreement and less random predictions..

Similarly, MCC provides a balanced measure even when class distributions are uneven:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The fact that, the transformer model has a high MCC further affirms the stability and strength of the model across every type of lesion.

The findings all show that performance of models tends to increase with the increase in the level of architectural complexity and representational capacity. Handcrafted characteristics are difficult to depict the visual complexity of dermoscopic images, and the interclass confusion is increased. CNNbased models are a much more powerful way of addressing this limitation, because they learn hierarchical features, and architecture improvements, including dualbranch connectivity, dense feature reuse and compound scaling, are also used to enhance discriminative power. The transformer based model has the

best performance because it can capture the long range dependencies as well as the global contextual relationships in the model courtesy of self attention. Transformers compare information on the whole image, unlike CNNs, which use localized receptive fields, which can represent lesions more holistically. This feature is especially beneficial to the separation of lesion types that look almost identical, as demonstrated by the near perfect confusion matrix and a high and consistent metric values.

Clinically, the improvements are very dramatic. Recall is high and the accuracy is balanced which minimises the risk of false negative, which is essential in the early diagnosis of malignant lesions like melanoma. The high metrics of agreement of the transformer based model also show that it is potentially applicable in real world clinical decision support. The model is promising as it has shown itself to be a reliable tool to aid dermatologists in the early and accurate diagnosis of skin cancer when compared to CNN based approaches in all aspects of evaluation.

5. CONCLUSION & FUTURE IMPROVEMENTS

This study outlined an end-to-end deep learning based approach for the early detection and multiclassification of skin cancer based on dermoscopic images. The work is a systematic evaluation and comparison of traditional machine learning, convolutional neural network (CNN), and transformer based models, all evaluated using a single experimental protocol, which enables a direct and systematic comparison of model performance at different levels of growing architecture complexity. The results show that the task of classifying skin lesions can be greatly improved using advanced deep learning techniques, especially for clinically realistic multiclass problems.

The experimental results indicate that there is a steady increase in the performance from the handcrafted feature-based SVM algorithms to CNN architectures such as ResNet18, DualBranch ResNet18, DenseNet121 and EfficientNetB0. All architectural improvements bring about better representation of features and less interclass confusion. The proposed transformer based model produces the best and strongest performance on all the evaluation metrics like accuracy, macroaveraged precision, recall, F1score, balanced accuracy and agreement based metrics. Its outstanding performance can be attributed to the self attention mechanism which can perform global contextual modeling and represent lesions in a more holistic way than localized convolutional operations. Despite these promising outcomes, there are a number of limitations. The research is based on a publicly accessible dataset, which, though varied, might not represent the full variation experienced in the clinical setting in actual practice, such as variations in imaging equipment, acquisition settings, and patient characteristics. Moreover, although, a good quantitative performance is shown, it should be validated on independent clinical data to evaluate the generalizability further.

The work in the future can be based on increasing the size of the dataset and covering more heterogeneous populations and rare types of lesions, introducing methods of explainable artificial intelligence to enhance the interpretability of the model, and introducing hybrid CNN transformer designs to gain additional performance. Moreover, prospective clinical studies and real time deployment scenarios should be explored to assess the practical impact of these models in the normal dermatological practice.

REFERENCES

Leiter, U., & Garbe, C. (2008). Epidemiology of melanoma and nonmelanoma skin cancer—The role of sunlight. *Advances in Experimental Medicine and Biology*, 624, 89–103. https://doi.org/10.1007/978-0-387-77574-6_8

Siegel, R. L., Miller, K. D., & Fuchs, H. E. (2023). Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, 73(1), 17–48. https://doi.org/10.3322/caac.21763

World Health Organization. (2022). *Skin cancers*. https://www.who.int/news-room/fact-sheets/detail/ultraviolet-(uv)-radiation

Narayanan, D. L., Saladi, R. N., & Fox, J. L. (2010). Ultraviolet radiation and skin cancer. *International Journal of Dermatology*, 49(9), 978–986. https://doi.org/10.1111/j.1365-4632.2010.04474.x

Ferris, L. K., et al. (2019). Artificial intelligence–based algorithm for the diagnosis of skin cancer in dermoscopic images. *JAMA Dermatology*, 155(1), 58–65. https://doi.org/10.1001/jamadermatol.2018.3265

Argenziano, G., et al. (2003). Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the Internet. *Journal of the American Academy of Dermatology*, 48(5), 679–693. https://doi.org/10.1067/mjd.2003.281

Rosendahl, C., Cameron, A., McColl, I., & Wilkinson, D. (2011). Dermoscopy in routine practice: "Chaos and clues". *Australasian Journal of Dermatology*, 53(1), 27–33. https://doi.org/10.1111/j.1440-0960.2011.00860.x

Tschandl, P., et al. (2019). Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26, 1229–1234. https://doi.org/10.1038/s41591-019-0552-0

Vestergaard, M. E., et al. (2008). Diagnostic accuracy of dermoscopy compared to naked eye examination of pigmented skin lesions. *British Journal of Dermatology*, 159(3), 669–676. https://doi.org/10.1111/j.1365-2133.2008.08713.x

Brinker, T. J., et al. (2019). Deep neural networks are superior to dermatologists in melanoma image classification. *European Journal of Cancer*, 119, 11–17. https://doi.org/10.1016/j.ejca.2019.05.032

Esteva, A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118. https://doi.org/10.1038/nature21056

Haenssle, H. A., et al. (2018). Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition. *European Journal of Cancer*, 120, 114–121. https://doi.org/10.1016/j.ejca.2019.07.006

Ali, M. A., et al. (2022). A comparative analysis of pre-trained models for skin lesion classification. *Biomedical Signal Processing and Control*, 71, 103157. https://doi.org/10.1016/j.bspc.2021.103157

He, K., et al. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on CVPR*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Huang, G., et al. (2017). Densely connected convolutional networks. *Proceedings of the IEEE Conference on CVPR*, 4700–4708. https://doi.org/10.1109/CVPR.2017.243

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning*, 6105–6114. https://arxiv.org/abs/1905.11946

Gessert, N., et al. (2020). Skin lesion classification using CNNs with patch-based training and hybrid features. *IEEE Journal of Biomedical and Health Informatics*, 24(2), 475–485. https://doi.org/10.1109/JBHI.2019.2919239

Codella, N. C., et al. (2018). Skin lesion analysis toward melanoma detection. *Nature*, 26(9), 1222–1228. https://doi.org/10.1038/s41591-019-0552-0

Combalia, M., et al. (2022). BCN20000: Dermoscopic lesion dataset and benchmark for multi-class classification. *Computer Methods and Programs in Biomedicine*, 198, 105770. https://doi.org/10.1016/j.cmpb.2020.105770

Carrera, C., & Marghoob, A. A. (2016). Discriminating nevi from melanoma. *Dermatologic Clinics*, 34(4), 395–409. https://doi.org/10.1016/j.det.2016.05.002

Wang, Q., et al. (2021). TransBTS: Multimodal brain tumor segmentation using transformer. *Medical Image Analysis*, 71, 102036. https://doi.org/10.1016/j.media.2021.102036

Dosovitskiy, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. https://arxiv.org/abs/2010.11929

Khan, S., et al. (2022). Transformers in vision: A survey. *ACM Computing Surveys*, 54(10), 1–41. https://doi.org/10.1145/3458306

Chen, J., et al. (2021). TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*. https://arxiv.org/abs/2102.04306

Raju, A. S. N., et al. (2025). XAI-SkinCADx: A six-stage explainable deep ensemble framework. *IEEE Transactions on Medical Imaging*. https://doi.org/10.1109/TMI.2025.11186509

Shoaib, B., et al. (2025). An explainable hybrid deep learning framework for precise skin lesion segmentation. *Frontiers in Medicine*, 12, 1681542. https://doi.org/10.3389/fmed.2025.1681542

Singh, A., et al. (2025). A hybrid CNN-transformer model for multi-class classification of dermatoscopic lesions. *ResearchGate Preprint*. https://doi.org/10.2139/ssrn.5339552

Cockayne, M. J., et al. (2025). DermFormer: Nested multi-modal vision transformers for robust skin cancer detection. *Pattern Analysis and Applications*, 28(1), 77–92. https://doi.org/10.1007/s10044-025-01572-0

Brinker, T. J., et al. (2022). Comparing the performance of balanced accuracy, MCC, and kappa statistics for AI classifiers. *BMC Medical Informatics and Decision Making*, 22(1), 58. https://doi.org/10.1186/s12911-022-01788-9

Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5, 180161. https://doi.org/10.1038/sdata.2018.161

Barata, C., et al. (2014). Improving dermoscopy image classification using color constancy. IEEE Journal of Biomedical and Health Informatics, 19(3), 1146–1152. https://doi.org/10.1109/JBHI.2014.2344612

Celebi, M. E., et al. (2015). Lesion border detection in dermoscopy images. Dermatology Research and Practice, 2015. https://doi.org/10.1155/2015/514785

Li, Y., et al. (2020). Dual-branch CNN for skin lesion classification. IEEE Access, 8, 128283–128292. https://doi.org/10.1109/ACCESS.2020.3008320

Abbas, Q., et al. (2013). Pattern classification of dermoscopic images. Skin Research and Technology, 19(1), 84–92. https://doi.org/10.1111/j.1600-0846.2012.00647.x

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. CVPR. https://doi.org/10.1109/CVPR.2016.90

Huang, G., et al. (2017). Densely connected convolutional networks. CVPR. https://doi.org/10.1109/CVPR.2017.243

Szegedy, C., et al. (2016). Inception-v4 and residual connections. AAAI.

Gessert, N., et al. (2020). Skin lesion classification using CNNs. Medical Image Analysis, 61, 101777. https://doi.org/10.1016/j.media.2020.101777

Brinker, T. J., et al. (2019). CNNs vs dermatologists. European Journal of Cancer, 119, 11–17. https://doi.org/10.1016/j.ejca.2019.05.032

Abdulameer, A., et al. (2025). Ensemble deep learning for skin cancer. International Journal of Electrical Engineering. https://doi.org/10.1016/j.ijee.2025.1034

Tanveer, J. (2025). Multi-class CNN for dermoscopy. ResearchGate.

Fiaz, M., et al. (2025). Explainable hybrid model. Frontiers in Medicine. https://doi.org/10.3389/fmed.2025.1681542

Ravula, J., et al. (2025). STViTDA-Net transformer. SSRN. https://doi.org/10.2139/ssrn.5339552

Brzozowski, K., et al. (2025). CNN trust evaluation. IEEE. https://doi.org/10.1109/TMI.2025.3056421

Tan, M., & Le, Q. (2019). EfficientNet. ICML. https://arxiv.org/abs/1905.11946

Howard, A. G., et al. (2017). MobileNets. arXiv preprint. https://arxiv.org/abs/1704.04861

Iandola, F. N., et al. (2016). SqueezeNet. arXiv preprint. https://arxiv.org/abs/1602.07360

Dosovitskiy, A., et al. (2020). Vision Transformers. arXiv preprint. https://arxiv.org/abs/2010.11929

Khan, S., et al. (2022). Transformers in vision: A survey. ACM Computing Surveys, 54(10), 1–41. https://doi.org/10.1145/3458306

Chen, J., et al. (2021). TransUNet. arXiv preprint. https://arxiv.org/abs/2102.04306

Cockayne, M. J., et al. (2025). DermFormer for skin cancer. Pattern Analysis and Applications. https://doi.org/10.1007/s10044-025-01572-0

Singh, A., et al. (2025). CNN-Transformer for skin cancer. ResearchGate. https://doi.org/10.13140/RG.2.2.33513.08802

Raju, A. S. N., et al. (2025). XAI-SkinCADx explainable model. IEEE. https://doi.org/10.1109/XAI-SkinCADx.2025.01

Shoaib, B., et al. (2025). Explainable hybrid DL. Frontiers in Medicine. https://doi.org/10.3389/fmed.2025.1681542

Tschandl, P., et al. (2018). HAM10000 dataset. Scientific Data, 5, 180161. https://doi.org/10.1038/sdata.2018.161

Combalia, M., et al. (2019). BCN20000 dataset. Data in Brief, 25, 104044. https://doi.org/10.1016/j.dib.2019.104044

Buda, M., et al. (2018). Class imbalance in CNNs. IEEE Access, 6, 42843–42852. https://doi.org/10.1109/ACCESS.2018.2851735

Chawla, N. V., et al. (2002). SMOTE. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953