



A Novel Vision Transformer Model for Skin Cancer Classification

Guang Yang¹ · Suhuai Luo¹ · Peter Greer¹

Accepted: 24 February 2023
© The Author(s) 2023

Abstract

Skin cancer can be fatal if it is found to be malignant. Modern diagnosis of skin cancer heavily relies on visual inspection through clinical screening, dermoscopy, or histopathological examinations. However, due to similarity among cancer types, it is usually challenging to identify the type of skin cancer, especially at its early stages. Deep learning techniques have been developed over the last few years and have achieved success in helping to improve the accuracy of diagnosis and classification. However, the latest deep learning algorithms still do not provide ideal classification accuracy. To further improve the performance of classification accuracy, this paper presents a novel method of classifying skin cancer in clinical skin images. The method consists of four blocks. First, class rebalancing is applied to the images of seven skin cancer types for better classification performance. Second, an image is preprocessed by being split into patches of the same size and then flattened into a series of tokens. Third, a transformer encoder is used to process the flattened patches. The transformer encoder consists of N identical layers with each layer containing two sublayers. Sublayer one is a multihead self-attention unit, and sublayer two is a fully connected feed-forward network unit. For each of the two sublayers, a normalization operation is applied to its input, and a residual connection of its input and its output is calculated. Finally, a classification block is implemented after the transformer encoder. The block consists of a flattened layer and a dense layer with batch normalization. Transfer learning is implemented to build the whole network, where the ImageNet dataset is used to pretrain the network and the HAM10000 dataset is used to fine-tune the network. Experiments have shown that the method has achieved a classification accuracy of 94.1%, outperforming the current state-of-the-art model IRv2 with soft attention on the same training and testing datasets. On the Edinburgh DERMOFIT dataset also, the method has better performance compared with baseline models.

✉ Guang Yang
guang.yang10@uon.edu.au

Suhuai Luo
<https://www.newcastle.edu.au/profile/suhuai-luo>

Peter Greer
<https://www.newcastle.edu.au/profile/peter-greer>

¹ School of Information and Physical Sciences, College of Engineering, Science and Environment, The University of Newcastle, University Dr., Callaghan, NSW 2308, Australia

Keywords Skin cancer classification · Deep learning · Transformer · Image processing · Neural networks

1 Introduction

Skin cancer is a common disease. Over five million new cases are diagnosed in the USA each year [1]. Australia has an even higher per capita incidence [2]. Among all the skin cancers, the most dangerous type is melanoma. It was reported that more than 190,000 new cases of melanoma were found in the USA in 2019 [3], contributing to the majority of deaths caused by all types of skin cancers. If diagnosed in the late stages, the survival rate for melanoma is poor [4]. Nevertheless, the survival rate can be as high as 97% [5] if it is found at early stages, which means early detection is vital. However, malignant skin cancer is not the primary form of skin cancer. Benign skin cancer makes up most of the cases. The similarity among the types makes the detection and classification of malignant skin cancer a challenging task. In current clinical practice, a high-resolution image solution, dermoscopy, is utilized by dermatologists to help diagnose skin cancer [6]. However, because a doctor usually requires extensive clinical visual experience to identify different types of skin cancers, the diagnosis result can be unreliable [7].

Whether it is a dermatologist or a machine learning algorithm, there has always been a problem accurately classifying some skin cancer types, especially malignant types. Therefore, deep learning techniques have been developed over the last few years to help improve the accuracy of diagnosis and the classification of skin cancers. In 2017, skin cancer classification by deep learning models and by dermatologists was compared [8]. The results showed that the latest deep learning methods outperformed dermatologists in accuracy and sensitivity of classification. Since then, more researchers have improved the performance of deep learning methods by using segmentation, data augmentation, and more advanced models [9]. However, for some malignant skin cancer types, such as melanoma, state-of-the-art algorithms still do not provide ideal classification accuracy due to similarity between types. In order to further improve the accuracy of deep learning algorithms in skin cancer classification, we launched research on improving the performance of deep learning algorithms in this field. Inspired by the work done in [12], we propose a new transformer network structure for skin cancer classification. We have compared our model's performance with the state-of-the-art models trained on the same dataset. Our experiment has shown that our transformer model achieved better accuracy and sensitivity than the non-transformer models. It can target the disease area on the dermoscopy image with less interference from the healthy area and noise, making it more suitable for skin lesion classification than the compared models.

2 Related Work

Traditional machine learning techniques were developed before deep learning was introduced into skin cancer classification. Researchers [33] used asymmetry, border, color, diameter (ABCD) as indicators to analyze skin lesions. Different algorithms are used for classification. A multilayer perceptron (MLP) network is used to classify features [33]. [34] enhanced images and removed noise by introducing a Gabor filter and geodesic active contours and then derived features by ABCD method. Logistic regression with artificial neural networks was also applied [35] to analyze the thickness of skin lesions. Additionally, there is a technique

to enhance the RGB channel of the image through the median filter [36] and then use the deformable model for segmentation. Support vector machines have also been used to classify features decomposed by ABCD rules [43] [37]. Similarly, K-means and K-nearest neighbor (KNN) are also applied to clustering and classification [38] [39]. CAD [40] [41] and 3D reconstruction [42] are also used for classification work.

In 2017, Esteva et al. [8] published their research comparing the performance of skin cancer classification by a pretrained GoogleNet Inception v3 model with the diagnoses of a group of dermatologists. Since then, an increasing number of researchers have contributed deep learning methods to this area, with many promising results published.

One method, designed by Nadipineni [11] in 2020, focused on the clinical image pre-processing step by conducting segmentation using U-Net and carrying out extensive data augmentation. They used ensemble technique and managed to achieve the best overall performance based on a series of fine-tuned convolutional neural network (CNN) models. The method achieved an accuracy of 92.6%, which was the state-of-the-art deep learning method in 2020.

In 2021, Datta et al. [10] implemented a deep learning method combining some of the best performing CNNs and soft attention layers, achieving an accuracy of 93.4%. The result outperforms Nadipineni's [11] approach on the HAM10000 dataset [13]. In their work, the data preprocessing technique removed duplicated data, and oversampling and undersampling of data was conducted to rebalance the number of images among the different types.

Moreover, the concept of vision transformer (ViT) was introduced in [12], from which our method is inspired. In the paper, the authors introduced a model consisting of multiple soft attention layers based on the transformer architecture previously used in natural language processing. By cutting the input image into small patches and transforming them into tokens, the model processed images similar to the way that sentences were processed in the transformers. After an extensive training dataset was pretrained on the model and fine-tuned on a smaller dataset, the ViT achieved good performance on some popular benchmark datasets, such as ImageNet.

After the ViT got researchers' attention because of its strong performance, there was some work in the field of skin cancer classification/segmentation based on the ViT model in 2022. In the work [30], researchers proposed a new method in the image feature embedding block of the original ViT model combined with a contrastive learning method. [32] conducted experiments addressing the bottleneck of the original ViT with their improved position encoding method. Both achieved comparable results compared to the previous best performer in skin cancer classification. Attempts with ViT have also been conducted in the skin lesion segmentation field [31]. By introducing a spatial pyramid pooling module into the original transformer's multihead attention, computation efficiency has been improved and better segmentation performance compared with baseline CNN models has been achieved.

Methodology.

The proposed method is based on the concept of the transformer, which was first introduced by Vaswani et al. [14] in 2017 for automated machine translation. It was introduced to solve sequence-to-sequence tasks with the ability to handle long-range dependencies. A transformer avoids recurrency but relies entirely on an attention mechanism to draw global relationships between input and output. It soon became the critical component of the best performing models, such as BERT [15] and GPT [16]. In the ViT model [12] introduced by Dosovitskiy, the original transformer's architecture was followed to utilize the scalability and efficiency of the design. The proposed method, though improved, is inspired by the ViT model. It is named a ViT for skin cancer detection (ViTfSCD). The method consists of four blocks as depicted in Fig. 1. The original dataset will first flow through the data augmentation and cancer-class

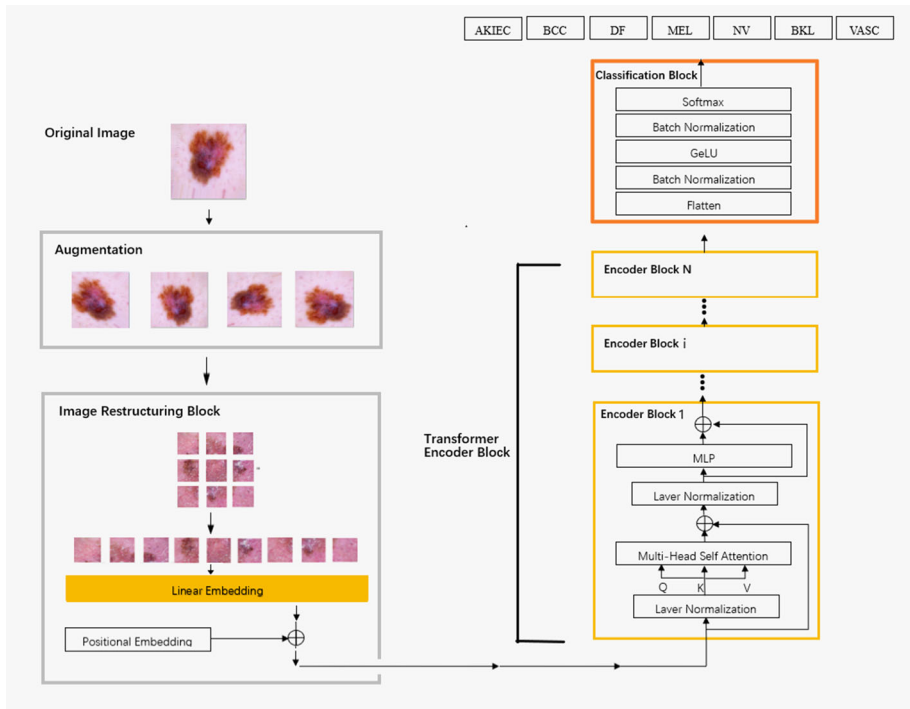


Fig. 1 The proposed vision transformer for skin cancer detection (ViTfSCD)

rebalancing block followed by the image restructuring block, transformer encoder block, and classification block. The details of these blocks are described below.

(a) The data augmentation and cancer-class rebalancing block

Dermoscopy is a method of skin lesion inspection using a device consisting of a high-resolution lens with a proper illumination setting. Dermoscopy images for skin lesions are becoming a popular source for artificial intelligence studies in recent research [8, 10, 11]. The dataset used in this study is the HAM10000 dataset [13] provided by ISIC. The dataset consists of 10,015 dermoscopic images, each with a size of 450 pixels by 600 pixels. It includes seven categories, including actinic keratosis intraepithelial carcinoma (AKIEC), basal cell carcinoma (BCC), benign keratosis (BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevi (NV), and vascular lesions (VASC), as illustrated in Fig. 2.

Class rebalancing is applied to the images of seven types of skin cancers in the training set for better classification performance. The class rebalancing is needed, as the dataset of

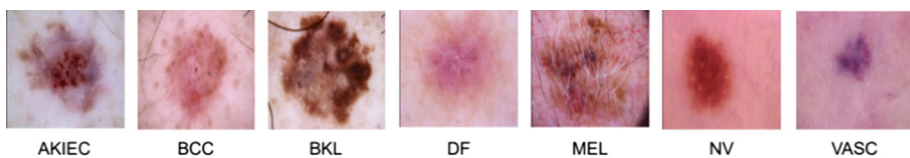


Fig. 2 Skin lesion examples in HAM10000 dataset

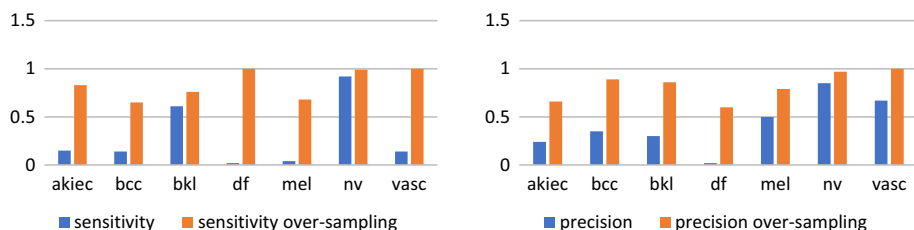


Fig. 3 Sensitivity/precision comparison for each type trained with/without oversampling

dermoscopy images contains an imbalanced number across different classes. The dataset is processed by oversampling and undersampling to ensure all the classes have an equal number of images. Oversampling is applied to the image class in the training dataset with fewer images by applying Keras data augmentation methods, including random rotations of 180 degrees, width and height shifts of 0.1 fractions, random zooming ranges of 0.1, and random flipping both horizontally and vertically. The accuracy of the same model (ViTfSCD-B) applied to the HAM10000 dataset was compared with and without oversampling. The result showed the overall accuracy can be improved from 70.46 to 91.4%. We can see from sensitivity figures in Fig. 3 that, before oversampling, types with far fewer samples were underfitting compared with types with more samples (such as BKL and NV). However, after oversampling, each class acquired more balanced training. Accordingly, we used the oversampled HAM10000 dataset for all our experiments. Undersampling is conducted by removing duplicated images of the same skin lesion. Finally, the pixel values of the images are normalized to a 0–1 range.

(b) Image restructuring block

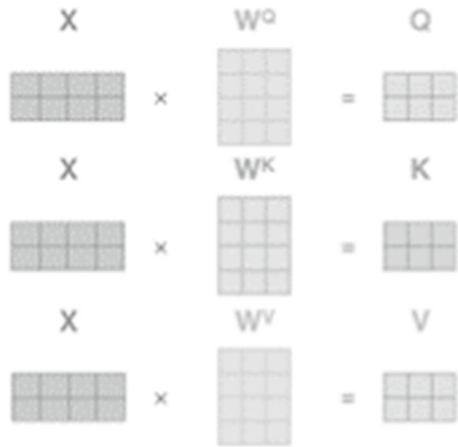
The image restructuring block consists of three processes. A 2D input image is first split into a sequence of patches of the same size. The number of the patches is the input sequence length that will be used in the transformer. Then, the patches go through a trainable linear embedding where a linear projection is performed to flatten the patches into a series of tokens with the same size, as transformers require constant latent vector size through all layers. Finally, a positional embedding is performed on the embedded tokens to retain the positional information of the patches where standard, learnable 1D position embeddings are utilized. The output of these processes will be the input to the following transformer encoder block.

(c) Transformer encoder block

The transformer encoder consists of N repeated layers, with each layer consisting of two sublayers. Sublayer one is a multihead self-attention layer [14]. Sublayer two is a fully connected feed-forward network. Two operations are applied on both sublayers, including normalization and residual connection. Normalization [21] is performed on every sublayer. Residual connections [22] between input and output are conducted for each sublayer. From the two operations, each sublayer has the output of $x + \text{sublayer}(\text{Norm}(x))$, where sublayer (x) is the function of the sublayer itself over input x .

To explain the concept of multihead self-attention, we need to look at an attention function first. An attention function can be described as mapping a query and a set of key-value pairs to an output. Attention is usually called scaled dot-product attention [14]. The input of the attention function is a group of vectors, including a query and a set of key-value pairs (as depicted in Fig. 4). The output is a weighted sum of the values where the weight is

Fig. 4 Queries (Q), keys (K), and values (V)



calculated by a compatibility function of the query with the related key. Attention is also called self-attention, since it relates different positions of a single input sequence to compute a representation of the sequence.

Assume the input consists of queries, keys of dimension d_k , and values of dimension d_v . The queries, keys, and values are packed together into matrices Q , K , and V . These three matrices are the result of multiplying the embedding of input vector X with each of the three weight matrices (W^Q, W^K, W^V) that are acquired via training. Then, a score is calculated by multiplying Q and K divided by the square root of the dimension of the key vectors, resulting in more stable gradients. The result is passed to a softmax function, of which the sum of all the outputs is 1. Then, the result is multiplied by the value vector V to create the attention for this input X . The attention is expressed as Eq. (1).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Multihead self-attention runs an attention function in parallel, allowing the model to jointly attend to information from different representation subspaces at different positions. The outputs of the operation are concatenated and then converted to linear units of the expected dimension [14]. This is expressed as Eq. (2).

$$\text{MutiHeadSelfAttention}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

where the projections are parameter matrices $W_i^Q \in R^{d_{model} \times d_k}$, $W_i^K \in R^{d_{model} \times d_k}$, $W_i^V \in R^{d_{model} \times d_v}$, and $W^O \in R^{hd_v \times d_{model}}$. Also, d_{model} is a constant value of the dimension.

The MLP sublayer is used for additional processing of the outputs. It consists of two fully connected layers and a GeLU nonlinearity layer [12].

(d) Classification block

The output of the set of N transformer encoders provides a good set of features for cancer classification. The classification block is specially designed for cancer classification. It consists of five layers, including a flatten layer, normalization layer 1, a dense layer, normalization layer

2, and a softmax layer. The flatten layer converts the multidimensional output of the transformer encoder into a 1D array. In order to recenter and rescale the distribution of the input to prevent vanishing or exploding gradients, the output of the flatten layer is batch normalized [21] and then inputted into the dense layer. The dense layer uses a GeLU activation function that was introduced initially in [24] and was proven to have higher performance than ReLU activation in a deep neural network. A second batch normalization is applied to the output of the dense layer. Finally, a softmax layer is applied to obtain the final classification result. Comparing the performances shows that our revised classification block performed slightly better than the original MLP head in ViT. The main difference was the batch normalization layers inserted after each dense layer in the classification block.

3 Experiments and Results

Four experiments were conducted to test the performance of four methods. The four methods included an Inception ResNet with Soft Attention (IRv2 + SA), a ResNet50 with Soft Attention (ResNet50 + SA) (provided by Datta et al. [10] with published code at GitHub), and two versions of the proposed ViTfSCD, including ViTfSCD-Large (ViTfSCD-L) and ViTfSCD-Base (ViTfSCD-B). Here the particular settings of the two ViTfSCD models are given in Table 3. This section presents the details of the experiments and the results of the proposed methods along with the compared methods.

(a) Data preparation

As mentioned above, the HAM10000 dataset was used for training and test data. The images were resized for the different methods: 384×384 for ViTfSCD, 299×299 for the IRv2 + SA, and 224×224 for the ResNet50 + SA, respectively.

The HAM10000 dataset includes seven classes of skin cancer, of which the number of images is summarized in Table 1. The dataset first underwent a preparation process that mainly included duplication removal, class rebalancing, and extended augmentation at the pixel level. The duplication removal removed duplicate images in the dataset. The class rebalancing process, as described in Sect. 3(a), aimed to allocate each class of the seven classes of training data the same number of images to achieve better classification performance. The extended augmentation at the pixel level added operations, such as random saturation, random contrast, and random brightness of pixels.

Table 1 Number of images for each skin cancer class in the HAM10000 dataset

Type of skin cancer	Number of images	Images in training set after oversampling
AKIEC	327	6,992
BCC	514	7,858
BKL	1,099	7,931
DF	115	6,876
MEL	1,113	7,903
NV	6,705	8,042
VASC	142	7,090
Total	10,015	52,692

Table 2 Parameters of two vision transformer for skin cancer detection (ViTfSCD) models

Model	Layers	Hidden size	MLP size	Heads	Parameters
ViTfSCD-base	12	768	3072	12	86 million
ViTfSCD-large	24	1024	5120	16	307 million

When Datta et al. [10] trained the dataset with different training-test splits, they found that 85% to 15% training and testing splits had better performance than 80% to 20% and 70% to 30% splits. In our experiment, 85% to 15% splits were used as well.

(b) Model training and fine-tuning

Both versions of the proposed ViTfSCD, i.e., ViTfSCD-L and ViTfSCD-B, have 16×16 patch size. However, they are different in their number of layers, hidden size, MLP size, and number of heads, as detailed in Table 2. Here, the layer is how many times the transformer encoder is repeated. The hidden size represents the token vector's length. The MLP size is the number of nodes of the first layer of the MLP section in each encoder block. The number of heads is the number of heads of the multihead attention. The parameters are the number of variables in the whole model that will learn during the training process.

The training of the ViTfSCD models started with pretrained weights from ImageNet for the original ViT model. According to [12], after applying weight from the pretrained model using a large dataset, ViT performance can improve when trained on a smaller dataset. A low learning rate for training was used to fine-tune the ViTfSCD models to the adopted datasets.

(c) Evaluation

In our experiment, accuracy, precision, sensitivity or recall, specificity, and F1 score were used to evaluate the performance of each model, of which the definitions are given below.

$$\begin{aligned}
 \text{Accuracy} &= \frac{\text{TruePositive} + \text{TrueNegative}}{\text{TotalNumber}} \\
 \text{Precision} &= \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \\
 \text{Sensitivity} &= \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \\
 \text{F1Score} &= \frac{2 * (\text{Precision} * \text{Sensitivity})}{\text{Precision} + \text{Sensitivity}} \\
 \text{Specificity} &= \frac{\text{TrueNegative}}{\text{TrueNegative} + \text{FalsePositive}}
 \end{aligned}$$

Accuracy is the ratio of the number of correctly predicted images divided by the total number of tested images; precision is the ratio of the number of correctly predicted positive images divided by the total number of predicted positive images in each type; sensitivity is the ratio of the number of correctly predicted positive images of the type divided by the total number of images of the type; F1 score is the weighted

Table 3 Comparison of each model’s precision, sensitivity, F1 score, and accuracy, where the values in **bold** represent the best evaluation parameters for each skin cancer type

Skin cancer type	Precision				Sensitivity				F1 score				Specificity			
	IRv2 + SA				ViTf SCD-B + SCD-L				IRv2 + SA				ViTf SCD-B + SCD-L			
	Res	Net	50 + SA	SA	Res	Net	50 + SA	SA	Res	Net	50 + SA	SA	Res	Net	50 + SA	SA
AKIEC	0.81	0.67	0.82	0.66	0.57	0.52	0.61	0.83	0.67	0.59	0.70	0.73	1.00	0.99	1.00	0.99
BCC	0.95	0.90	1.00	0.89	0.73	0.69	0.54	0.65	0.83	0.78	0.70	0.76	1.00	1.00	1.00	1.00
BKL	0.73	0.72	0.81	0.86	0.79	0.70	0.65	0.76	0.76	0.71	0.72	0.81	0.98	0.98	0.99	0.99
DF	0.62	0.62	0.67	0.60	0.83	0.83	1.00	1.00	0.71	0.71	0.80	0.75	1.00	1.00	1.00	1.00
MEL	0.70	0.62	0.60	0.79	0.47	0.59	0.53	0.68	0.56	0.61	0.56	0.73	0.99	0.99	0.99	0.99
NV	0.95	0.96	0.94	0.97	0.97	0.98	0.98	0.99	0.96	0.97	0.96	0.98	0.80	0.84	0.76	0.88
VASC	0.91	0.83	0.83	1.00	1.00	1.00	1.00	1.00	0.95	0.91	0.91	1.00	1.00	1.00	1.00	1.00

average of sensitivity and precision, considered a better indicator of the classifier's performance than the regular accuracy. Specificity is the ratio of the number of correctly predicted negative images of the type divided by the total number of images not in the type.

Experiments were conducted on the classification performance of the four methods, including IRv2 + SA, ResNet50 + SA, ViTfSCD-L, and ViTfSCD-B. The results are given in Table 3. Note: For accuracy, only overall accuracy is given, and the individual accuracy for each skin cancer type is not given. The individual accuracy for each skin cancer type is not given since the individual accuracy values do not reflect the classification performance when the case numbers of the seven cancer types differ greatly in the test dataset.

In the table, all the best values of each performance for each cancer type are marked in bold. It can be seen from the table that the ViTfSCD-L model had the highest precision, sensitivity, and F1 scores. Moreover, the overall accuracy of ViTfSCD-L was the highest (94.1%), and the overall accuracy of ViTfSCD-B was 91.4%. Figure 5 below shows the comparison of precision, sensitivity, and F1 score across seven types of skin cancer and four tested models.

(d) Comparison of overall accuracy across models

We compared our result with the state-of-the-art methods of the HAM10000 dataset in Table 5. Two earlier methods were also referenced for comparison. Since only the accuracy was available from the two papers, only accuracy is compared here. The first was the work of Nielsen et al. [17] in 2020, in which they built an ensemble of multi-resolution Efficient-Nets, achieving an accuracy of 92.6% from the HAM10000 dataset. The second was the semi-supervised medical image classification method conducted by Liu et al. [18] in 2021, achieving an accuracy of 92.5% with the same dataset. These two works were based on data splitting and data augmentation methods with the HAM10000 dataset, while the results of the remaining six methods were based on the same training and testing dataset but our data augmentation method. In addition, the results achieved with ResNet50, ResNet50 + SA, IRv2, and IRv2 + SA in their original papers are shown along with the results that we achieved in our experiments. Since the research [10] shows evidence that Soft Attention may improve baseline models (ResNet, IRv2), we did not test the baseline models again in our work. However, we conducted the experiments using the original ViT models (ViT-Base 16 and ViT-Large 16) on the same dataset with the same approach, and the results show that our ViTfSCD-B and ViTfSCD-L models outperformed the original ViT models.

It can be seen from Table 4 that our fine-tuned ViTfSCD-L model outperformed other models by 1–2% in overall accuracy.

We also conducted the experiment using the Edinburgh DERMOFIT dataset [28], which is a clinical skin lesion dataset with 10 classes, including basal cell carcinoma, squamous cell carcinoma, intraepithelial carcinoma, malignant melanoma, melanocytic nevus, seborrheic keratosis, pyogenic granuloma, hemangioma, and dermatofibroma. The dataset contains 1300 images taken with a digital camera under similar settings. In earlier research [29], traditional methods were compared and showed that ResNet50 performed better than decision tree and KNN. In our experiment, we compared ResNet50, IRv2, and our ViTfSCD-B on the same training/test split and augmentation method. The results in Table 5 indicate that ViTfSCD-B outperformed other models.

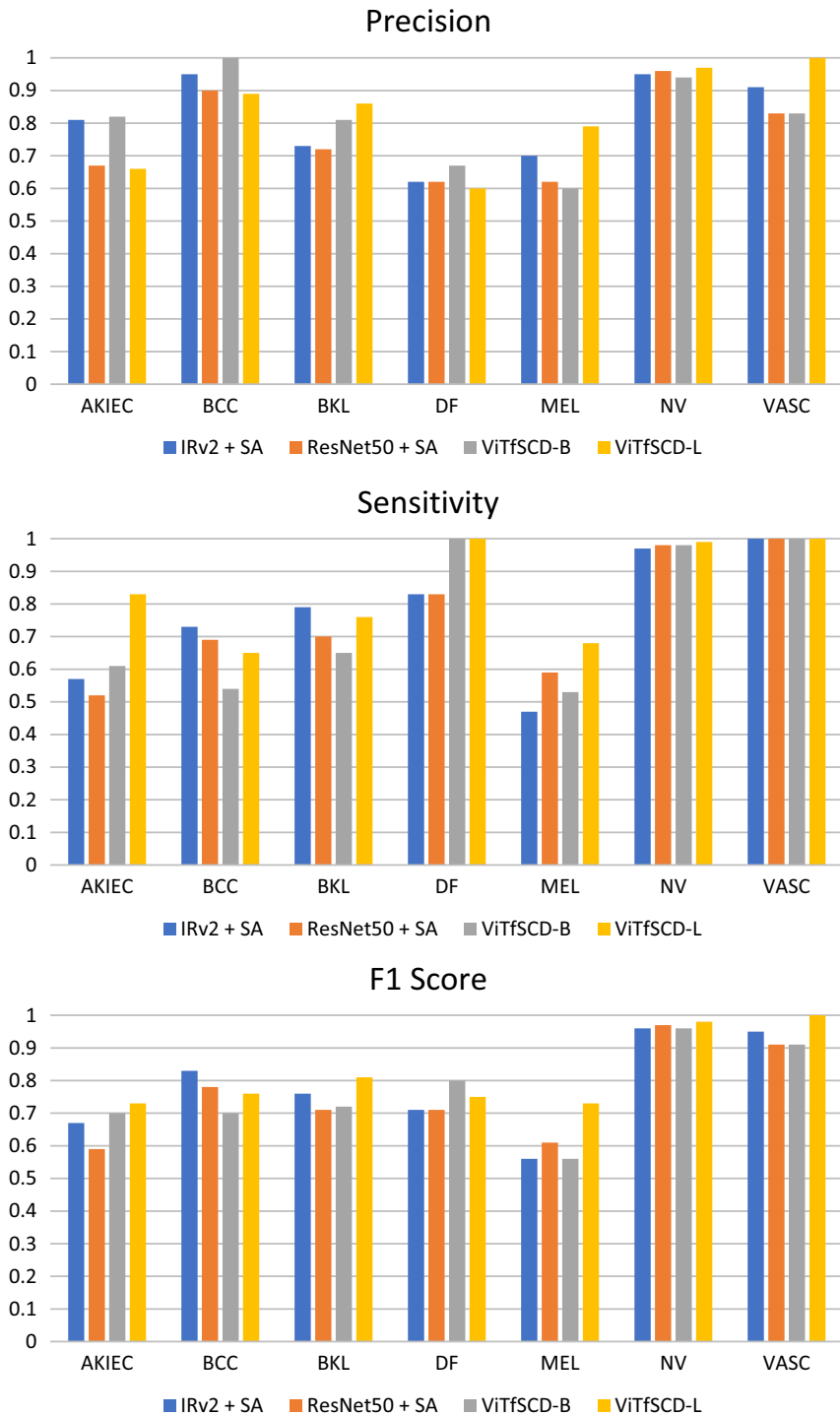


Fig. 5 Comparison of precision, sensitivity, and F1 score across seven types of skin cancer and four tested models

Table 4 Comparison of overall accuracy across models with the HAM10000 dataset

Methods	Year	Overall accuracy in their paper	Overall accuracy in our experiments
Loss balancing and ensemble [17]	2020	92.6%	–
Semi-supervised [18]	2021	92.54%	–
ResNet50 [10]	2021	90.5%	
IRv2 [10]	2021	91.2%	
ResNet50 with Soft Attention [10]	2021	91.5%	91.5%
Inception ResNet with Soft Attention [10]	2021	93.4%	91.9%
Vision transformer (ViT)-Base16	2021	–	91.1%
ViT-Large16	2021	–	93.7%
ViT for skin cancer detection-Base	2022	–	91.4%
ViT for skin cancer detection-Large	2022	–	94.1%

Table 5 Comparison of overall accuracy across models with the DERMOFIT dataset

Methods	Year	Overall accuracy in their paper	Overall accuracy in our experiments
Decision tree [29]	2020	78.1%	–
Flat ResNet50 [29]	2020	78.7%	–
ResNet50	2022	–	71.2%
Inception ResNet (IRv2)	2022	–	71.9%
IRv2 + Soft Attention	2022	–	75.0%
Vision transformer for skin cancer detection-Base	2022	–	80.5%

4 Discussion and Future Work

In this section, first, the effect of an attention layer is examined using the attention map feature of the attention layer. Then, a performance analysis of transformer models on different sizes is presented. Finally, future improvements are proposed.

(a) Attention map

One of the essential features in ViTfSCD is the attention map, which is the output of the attention layers in the transformer encoder block. Figure 6 shows an original image (left side) and its attention map (right side). In the attention map, the skin lesion areas are brighter, whereas the areas with less relevance to the illness are darker.

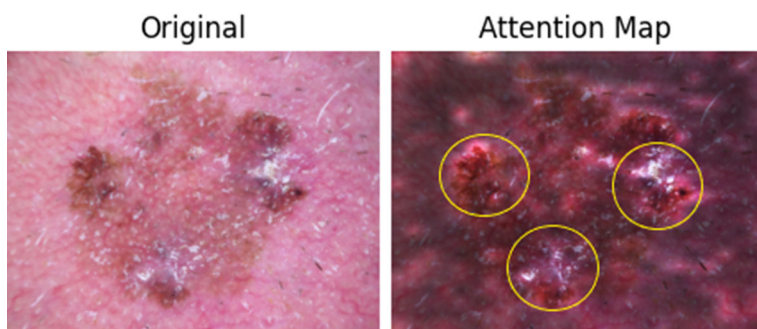


Fig. 6 An original image (left) and its attention map (right). In the attention map, the diseased areas are brighter

The attention map gives smaller weights to the irrelevant parts of the image. The irrelevant parts have less influence on the output or classification result of the model. The smaller the weights for the irrelevant parts are, the higher accuracy the classification can achieve (Fig. 6).

When comparing this feature to the traditional visual explanation of algorithms in deep learning, Datta et al. [10] found that attention generally performs better than methods such as Grad-CAM heat map [10]. The attention map shows that the diseased area is more relevant to the output of the entire model with higher weights. As the diseased areas highlighted by the attention map are assigned higher weights than the non-diseased areas, the highlighted areas have more influence on the output of the model, resulting in more accurate detection.

(b) The performance of transformer models of different size

In the work of [12], ViTs were classified into three types: ViT-Base, ViT-Large, and ViT-Huge according to the number of layers, hidden size, MLP size, and number of heads. Generally, larger models have larger layers and sizes with more parameters. ViT-Base has 86 million parameters, ViT-Large has 307 million parameters, and ViT-Huge has 632 million parameters. Following the model type is patch size. For example, ViT-L/16 stands for ViT-Large with a patch size of 16×16 , and ViT-B/32 stands for ViT-Base with a patch size of 32×32 . The smaller the patch size, the more expensive the computation, and the better performance it can achieve [12].

In 2021, an even bigger model, ViT-G/14 was built [20]. It further improved the accuracy of the ImageNet classification task to the state-of-the-art level of 90.45% from the 88.04% achieved by ViT-H/14.

(c) Limitations of current ViT models

Although the ViT model brings performance gains, we should also note its limitations. First, the volume of the transformer model increases significantly compared to various baseline models. For example, ResNet50 has 25 million parameters, IRv2 has 56 million parameters, and ViT-Large16 and ViTfSCD-L, which achieved better performance in our experiments, have 307 million parameters [12]. The more powerful model ViT-H/14 has 632 million parameters. The size of the model can significantly increase the cost of training and deployment.

Second, data efficiency has been an open research problem since the birth of transformer. ViT models normally require extensive pretraining to achieve a boost in performance. For example, ViT-Huge and ViT-G/14 both achieved outstanding performance with the benchmark ImageNet dataset after being pretrained on a 300 million image dataset—JFT-300M [19]. Large data requirements can be a bottleneck of the ViT models since the data limitation exists in many domains.

(d) Frontier and open areas

Model generalization is a key topic in skin cancer classification work. Most of the research so far is based on dermatological images acquired by medical imaging equipment. When these models are applied to images obtained by other devices, such as mobile devices or regular cameras, there tends to be a noticeable drop in performance. Transfer learning is one possible solution, but only if there is a reliable training dataset in the new domain, which is often missing.

Model robustness is another topic in skin cancer classification. Various major training sets currently provide high-quality and high-resolution images, which results in models that are less tolerant of noise and disturbances. One example is images collected by mobile phones, which tend to have different lighting conditions, perspectives, backgrounds, and resolutions. This also leads to a significant drop in model performance.

The efficiency of the model is also important. A large number of recent studies have improved model performance but have also greatly increased model size and training difficulty. The size of the model determines whether the model can be easily adapted to various devices with different computing and storage capabilities. The difficulty of training directly increases the cost of training.

Similarly, the limitations of the training set also limit the performance of the model. The imbalance in the amount of data across categories is an obvious problem. In some datasets, common skin cancer categories, such as BCC and SCC, occupy a large amount of data, while other less common but more dangerous skin cancer categories lack data, and some categories are even missing. One of the results is overfitting. Current solutions include generating samples through algorithms or augmentation and assigning different class weights to the loss function.

5 Conclusion

This paper presents a novel method of skin cancer classification using clinical skin images. The method consists of four blocks, including skin cancer class rebalancing for better classification performance, image splitting and tokenization for input preparation, transformer encoding to extract key features for classification, and final skin cancer classification. Transfer learning is implemented to build the whole network, where the ImageNet dataset is used to pretrain the network and the HAM10000 dataset is used to fine-tune the network. Experiments have shown that the method achieves a classification accuracy of 94.1%, outperforming the current state-of-the-art model, IRv2 + SA, on the same training and testing datasets. The experiment has also shown that the attention map feature brought by the attention layers further boosts the model's performance, making it a novel classification model for skin cancer.

6 Future Work

Our ViTfSCD model can be further investigated and improved in two aspects, including larger size and transfer learning. As scaling up transformer models has been shown to improve accuracy, there is therefore a strong possibility that applying larger transformer models to skin cancer classification tasks can also enhance the performance of classification. More studies will be conducted to seek better performance from larger ViTfSCD models. In developing ViTfSCD models, we implemented transfer learning by using the pretrained weights for ViT

from ImageNet-21 k published by Google [12]. Future work can focus on developing more models by utilizing pretrained weights on a larger dataset, such as JFT-300 M [19].

Acknowledgements This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' Contributions Guang Yang prepared the data, conducted the experiment, and drafted the manuscript. Suhui Luo reviewed the draft and enhanced the introduction, methodology, and conclusion sections. Peter Greer reviewed the manuscript and made suggestions. All authors reviewed the final manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Declarations

Competing interests None declared.

Ethical approval Not required.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Siegel RL, Naishadham D, Jemal A (2012) Cancer statistics. *CA* 62(1):10–29. <https://doi.org/10.3322/caac.20138>
2. Australian Bureau of Statistics (2019) Causes of Death, Australia [Internet]. ABS, Canberra. Accessed 2022 Nov 1. <https://www.abs.gov.au/statistics/health/causes-death/causes-death-australia/2019>.
3. Street W (2019) Cancer Facts & Figures. American Cancer Society, Atlanta, GA. <http://cancerstatisticscenter.cancer.org>. Accessed 2022 Nov 1.
4. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA* 68(6):394–424
5. Siegel RL, Miller KD, Jemal A (2019) Cancer statistics. *CA* 69(1):7–3
6. Vestergaard ME, Macaskill PH, Holt PE, Menzies SW (2008) Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. *Br J Dermatol* 159(3):669–676
7. Menzies SW, Bischof L, Talbot H, Gutenev A, Avramidis M, Wong L, Lo SK, Mackellar G, Skladnev V, McCarthy W, Kelly J (2005) The performance of SolarScan: an automated dermoscopy image analysis instrument for the diagnosis of primary melanoma. *Arch Dermatol* 141(11):1388–1396
8. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639):115–118
9. Adeyinka AA, Viriri S (2018) Skin lesion images segmentation: a survey of the state-of-the-art. In: International conference on mining intelligence and knowledge exploration. Springer, Cham, pp. 321–330
10. Datta SK, Shaikh MA, Srihari SN (2021) Soft Attention Improves Skin Cancer Classification Performance. In: Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data. Springer, Cham, pp 13–23
11. Nadipineni H (2020) Method to classify skin lesions using dermoscopic images. arXiv preprint [arXiv:2008.09418](https://arxiv.org/abs/2008.09418). 2020 Aug 21.

12. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929). 2020 Oct 22
13. Tschandl P, Rosendahl C, Kittler H (2018) The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data* 5(1):1–9
14. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: *Advances in neural information processing systems*. 2017, 30
15. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805). 2018 Oct 11.
16. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. *OpenAI blog* 1(8):9
17. Gessert N, Nielsen M, Shaikh M, Werner R, Schlaefer A (2020) Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. *MethodsX* 1(7):100864
18. Liu Q, Yu L, Luo L, Dou Q, Heng PA (2020) Semi-supervised medical image classification with relation-driven self-ensembling model. *IEEE Trans Med Imaging* 39(11):3429–3440
19. Sun C, Shrivastava A, Singh S, Gupta A (2017) Revisiting unreasonable effectiveness of data in deep learning era. In: *Proceedings of the IEEE international conference on computer vision 2017*, pp 843–852
20. Zhai X, Kolesnikov A, Houlsby N, Beyer L (2022) Scaling vision transformers. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2022*, pp 12104–12113
21. Ba JL, Kiros JR, Hinton GE. Layer normalization. arXiv preprint [arXiv:1607.06450](https://arxiv.org/abs/1607.06450). 2016 Jul 21.
22. 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 2016*, pp 770–778
23. Tomita N, Abdollahi B, Wei J, Ren B, Suriawinata A, Hassanpour S (2019) Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. *JAMA Netw Open* 2(11):e1914645
24. Hendrycks D, Gimpel K. Gaussian error linear units (gelus). arXiv preprint [arXiv:1606.08415](https://arxiv.org/abs/1606.08415). 2016 Jun 27.
25. Melas-Kyriazi L. Do you even need attention? a stack of feed-forward layers does surprisingly well on imagenet. arXiv preprint [arXiv:2105.02723](https://arxiv.org/abs/2105.02723). 2021 May 6.
26. Tolstikhin IO, Houlsby N, Kolesnikov A, Beyer L, Zhai X, Unterthiner T, Yung J, Steiner A, Keysers D, Uszkoreit J, Lucic M (2021) Mlp-mixer: an all-mlp architecture for vision. *Adv Neural Inf Process Syst* 6(34):24261–24272
27. Touvron H, Bojanowski P, Caron M, Cord M, El-Nouby A, Grave E, Izacard G, Joulin A, Synnaeve G, Verbeek J, Jégou H (2022) Resmlp: feedforward networks for image classification with data-efficient training. *IEEE Trans Pattern Anal Mach Intell*
28. Ballerini L, Fisher RB, Aldridge B, Rees J. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In: *Color medical image analysis 2013*. Springer, Dordrecht, pp 63–86
29. Fisher RB, Rees J, Bertrand A. Classification of ten skin lesion classes: Hierarchical knn versus deep net. In: *Annual conference on medical image understanding and analysis 2019 Jul 24*. Springer, Cham, pp 86–98
30. Xin C, Liu Z, Zhao K, Miao L, Ma Y, Zhu X, Zhou Q, Wang S, Li L, Yang F, Xu S (2022) An improved transformer network for skin cancer classification. *Comput Biol Med* 1(149):105939
31. He X, Tan EL, Bi H, Zhang X, Zhao S, Lei B (2022) Fully transformer network for skin lesion analysis. *Med Image Anal* 1(77):102357
32. Nakai K, Chen YW, Han XH (2022) Enhanced deep bottleneck transformer model for skin lesion classification. *Biomed Signal Process Control* 1(78):103997
33. Alencar FE, Lopes DC, Neto FM (2016) Development of a system classification of images dermoscopic for mobile devices. *IEEE Latin Am Trans* 14(1):325–330
34. Kasmi R, Mokrani K (2016) Classification of malignant melanoma and benign skin lesions: implementation of automatic ABCD rule. *IET Image Proc* 10(6):448–455
35. Sáez A, Sánchez-Monedero J, Gutiérrez PA, Hervás-Martínez C (2015) Machine learning methods for binary and multiclass classification of melanoma thickness from dermoscopic images. *IEEE Trans Med Imaging* 35(4):1036–1045
36. Ma Z, Tavares JM (2015) A novel approach to segment skin lesions in dermoscopic images based on a deformable model. *IEEE J Biomed Health Inform* 20(2):615–623
37. Pathan S, Prabhu KG, Siddalingaswamy PC (2019) Automated detection of melanocytes related pigmented skin lesions: a clinical framework. *Biomed Signal Process Control* 1(51):59–72

38. Kharazmi P, AlJasser MI, Lui H, Wang ZJ, Lee TK (2016) Automated detection and segmentation of vascular structures of skin lesions seen in Dermoscopy, with an application to basal cell carcinoma classification. *IEEE J Biomed Health Inform* 21(6):1675–1684
39. Dalila F, Zohra A, Reda K, Hocine C (2017) Segmentation and classification of melanoma and benign skin lesions. *Optik* 1(140):749–761
40. Noroozi N, Zakerolhosseini A (2016) Computer assisted diagnosis of basal cell carcinoma using Z-transform features. *J Vis Commun Image Represent* 1(40):128–148
41. Zakeri A, Hokmabadi A (2018) Improvement in the diagnosis of melanoma and dysplastic lesions by introducing ABCD-PDT features and a hybrid classifier. *Biocybern Biomed Eng* 38(3):456–466
42. Satheesha TY, Satyanarayana D, Prasad MG, Dhruve KD (2017) Melanoma is skin deep: a 3D reconstruction technique for computerized dermoscopic skin lesion classification. *IEEE J Transl Eng Health Med* 16(5):1–7
43. Oliveira RB, Marranghello N, Pereira AS, Tavares JM (2016) A computational approach for detecting pigmented skin lesions in macroscopic images. *Expert Syst Appl* 1(61):53–63

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.