



# Exploring the Correlation Between Deep Learned and Clinical Features in Melanoma Detection

Tamal Chowdhury<sup>1,2</sup>, Angad R.S. Bajwa<sup>1,3</sup>, Tapabrata Chakraborti<sup>4</sup>(✉), Jens Rittscher<sup>4</sup>, and Umapada Pal<sup>1</sup>

<sup>1</sup> CVPR Unit, Indian Statistical Institute, Kolkata, India

<sup>2</sup> National Institute of Technology (NIT), Durgapur, India

<sup>3</sup> National Institute of Technology (NIT), Tiruchirappalli, India

<sup>4</sup> IBME/BDI, Dept. of Engineering Science, University of Oxford, Oxford, UK  
[tapabrata.chakraborty@eng.ox.ac.uk](mailto:tapabrata.chakraborty@eng.ox.ac.uk)

**Abstract.** Despite the recent success of deep learning methods in automated medical image analysis tasks, their acceptance in the medical community is still questionable due to the lack of explainability in their decision-making process. The highly opaque feature learning process of deep models makes it difficult to rationalize their behavior and exploit the potential bottlenecks. Hence it is crucial to verify whether these deep features correlate with the clinical features, and whether their decision-making process can be backed by conventional medical knowledge. In this work, we attempt to bridge this gap by closely examining how the raw pixel-based neural architectures associate with the clinical feature based learning algorithms at both the decision level as well as feature level. We have adopted skin lesion classification as the test case and present the insight obtained in this pilot study. Three broad kinds of raw pixel-based learning algorithms based on convolution, spatial self-attention and attention as activation were analyzed and compared with the ABCD skin lesion clinical features based learning algorithms, with qualitative and quantitative interpretations.

**Keywords:** Explainable artificial intelligence · Melanoma classification · Digital dermatoscopy · Attention mechanisms · Deep machine learning

## 1 Introduction

Among the several variants of skin lesion diseases, melanoma is the condition that puts patients' lives at risk because of its highest mortality rate, extensive class variations and complex early stage diagnosis and treatment protocol. Early detection of this cancer is linked to improved overall survival and patient health.

T. Chakraborti is funded by EPSRC SeeBiByte and UKRI DART programmes.

T. Chowdhury and A. R. S. Bajwa—First authors with equal contributions.

© Springer Nature Switzerland AG 2021

B. W. Papież et al. (Eds.): MIUA 2021, LNCS 12722, pp. 3–17, 2021.

[https://doi.org/10.1007/978-3-030-80432-9\\_1](https://doi.org/10.1007/978-3-030-80432-9_1)

Manual identification and distinction of melanoma and its variants can be a challenging task and demands proper skill, expertise and experience of trained professionals. Dermatologists consider a standard set of features (popularly known as the ABCDE features) that takes into consideration the size, border irregularity, colour variation for distinguishing malignant and benign tumours. With proper segmentation boundaries these features can be extracted from images and used as inputs to machine learning algorithms for classification purposes. Also, with the recent advancements of deep learning, Convolutional neural networks (CNNs) [1] are able to differentiate the discriminative features using raw image pixels only [26]. But the decision making process of these complex networks can be opaque. Several approaches have been proposed to identify the image regions that a CNN focuses on during its decision-making process [28–30]. Van Molle et al. [3] tried to visualize the CNN learned features at the last layers and identified where these networks look for discriminative features. Young et al. [4] did a similar work towards the interpretability of these networks using GradCAM [5] and kernel SHAP [6] to show how unpredictable these models can be regarding feature selection even when displaying similar performance measures. Both these works demonstrated how pixel-based models can be misguided towards image saliency and focus on undesirable regions like skin hairs, scale marks etc. Also, attention guided CNNs were used [7, 8] to solve the issue of feature localization. Though these works provide a comprehensive insight to *where* these CNNs look for unique elements in an image they are not sufficient to unveil *what* exactly these models look for and more importantly, if there is any kind of correlation with their extracted sets of features and those sought by dermatologists (the *why* question). As the consequences of a false negative can be quite severe for such diagnostic problems, it is of utmost importance to determine if the rules learned by these deep neural networks for decision making in such potential life-threatening scenarios can be backed by medical science. In this paper we have tried to address this issue by experimenting with both handcrafted ABCD features and raw pixel based features learned by a deep learning models, along with exploring if there is any correlation present between them.

## 2 Dataset and Methodology

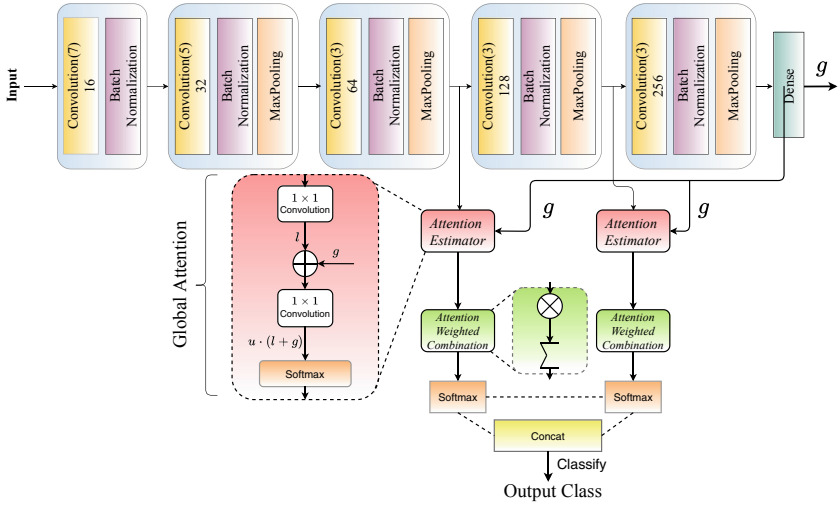
### 2.1 Dataset: Description and Pre-processing

**HAM10000** dataset [9] a benchmark data set for skin lesion classification, is used in this study. The dataset contains a total of 10015 dermoscopic images of dimensions  $3 \times 450 \times 600$  distributed over 7 classes namely: melanoma (Mel, 1113 samples), melanocytic nevi (NV, 6705 samples), basal cell carcinoma (BCC, 514 samples), actinic keratosis and intraepithelial carcinoma (AKIEC, 327 samples), benign keratosis (BKL, 1099 samples), dermatofibroma (DF, 115 samples) and vascular lesions (VASC, 142 samples).

**Pre-processing** steps are carried out to remove the artifacts. First, the images are center cropped to extract the main lesion region and separate out

the natural skin area, scale marks and shadows present due to the imaging apparatus [27]. Further, to remove the body hair and remaining scale marks, a local adaptive thresholding method is used where the threshold value of a pixel is determined by the range of intensities in its local neighbourhood. Finally, the images were enhanced using CLAHE [10] technique, and scaled using maximum pixel value. The entire dataset is divided in a 80 : 10 : 10 ratio as the training, validation and test set, respectively.

## 2.2 Deep Architectures



**Fig. 1.** An overview of the baseline CNN architecture along with the global attention modules attached to the last two convolutional blocks.

**Baseline CNN:** First, we designed a simple convolutional neural network with 5 convolutional blocks that serves as the baseline for other deep learning models used in this paper. Each convolutional block further consists of a convolutional layer followed by ReLU activation, max pooling (except for the first block) and batch normalization layers. Dropout layers with a dropout probability of 0.2 were used after the convolutional layers of the last two blocks to reduce overfitting. The convolutional blocks are then followed by global average pooling (GAP) [17] suitable for fine-grain classification problems and a softmax based classification layer. We used convolutional kernels with spatial extent 7, 5, 3, 3, 3 for consecutive convolutional blocks with 16, 32, 64, 128 and 256 feature maps, respectively.

**CNN with Global Attention:** Considering the importance for the network to focus on clinically relevant features, we further test the network by adding global attention modules, proposed by Jetley et al. [18] on top of the last two convolutional blocks of our baseline CNN model. The resulting network is presented in Fig. 1, which is end-to-end trainable. This method exploits the universality between local and global feature descriptors to highlight important features of an input.

First a *compatibility* score ( $c_i^s$ ) is calculated using the local feature vector  $l_i^s$  and the global feature vector  $g$  as:

$$c_i^s = u(l_i^s + g) \quad (1)$$

Where,  $l_i^s$  represents the  $i_{th}$  feature map of  $s_{th}$  convolutional layer. Here  $i \in \{1, 2, \dots, n\}$  and  $s \in \{1, 2, \dots, S\}$  ( $n$  = number of feature maps and  $S$  = total number of layers in the network).  $u$  is the weight vector learning the universal feature sets for the relevant task.  $1 \times 1$  convolutions are used to change the dimensionality of  $l^s$  to make it compatible for addition with  $g$ . Next, the attention weights  $a$  are calculated from the compatibility scores  $c$  by simply applying a softmax function function as:

$$a_i^s = \frac{\exp(c_i^s)}{\sum_{k=1}^n \exp(c_k^s)} \quad (2)$$

These two operations sum up as the *attention estimator*. The final output of the attention mechanism for each block  $s$  is then calculated as:

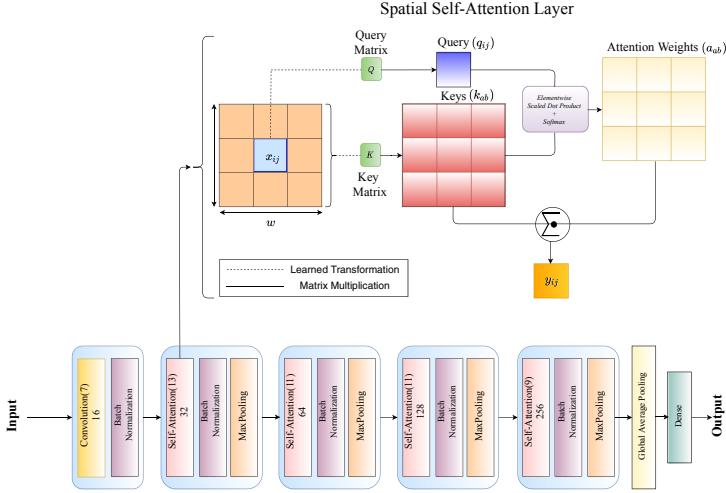
$$g_a^s = \sum_{i=1}^n a_i^s \cdot l_i^s \quad (3)$$

Two such  $g_a^s$ s are concatenated as shown and a dense layer is added on top of it, to make the final prediction.

**Spatial Self-attention Model:** Inspired by the enormous success achieved by the transformer networks [19] in the field of natural language processing (NLP), Ramachandran et al. [20] proposed a classification framework in spatial domain entirely based on self-attention. The paper showed state-of-the-art performance on multiple popular image datasets questioning the need for convolution in vision tasks. Like convolution, the fundamental goal of self-attention also is to capture the spatial dependencies of a pixel with its neighbourhood. It does so by calculating a similarity score between a query pixel and a set of key pixels in the neighbourhood with some spatial context.

Here, we have modified the self-attention from the original work. In case of self-attention, the local neighbourhood  $\mathcal{N}_k$  is denoted as the *memory block*. In contrast to the global attention modules, here the attention is calculated over a local region, which makes it flexible for using at any network depth without

causing excessive computational burden. Two different matrices, query ( $q_{ij}$ ) and keys ( $k_k$ ), are calculated from  $x_{ij}$  and  $x_{ab} \in \mathcal{N}_k$ , respectively by means of linear transformations as shown in Fig. 2. Here,  $q_{ij} = Qx_{ij}$  and  $k_k = Kx_{ab}$  where  $Q, K \in \mathbb{R}^{d_{out} \times d_{in}}$  are the query and key matrices respectively and formulated as the model parameters. Intuitively, the query represents the information (pixel) *to be matched* with a look up table containing the addresses and numerical values of a set of information represented by the keys.



**Fig. 2.** An overview of the proposed self-attention model. Query ( $q_{ij}$ ) and Keys ( $k_k$ ) are calculated from  $x_{ij}$  and its neighbourhood  $x_{ab}$  by their linear transformations using  $Q$  and  $K$  matrices respectively.

In the original work [20], proposing the self-attention layer in spatial domain, a separate value matrix  $V$  is taken to calculate the values  $v$  which is a linear projection of the original information. Technically, in our case, keys are essentially the same thing, with the keys containing extra positional information that has been added explicitly. So, we've discarded  $v$  entirely and used  $k$  only for calculating the attention weights as well as representing the original information; that reduces the total number of model parameters. In practice, the input is divided into several parts along the depth (feature maps) and multiple convolution kernels. Multiple such query-key matrix pairs known as *heads* are used to learn distinct features from an input. Unlike [20], the single headed *normalized* attention scores in the neighbourhood  $\mathcal{N}_k$  are calculated as the scaled dot product of queries and keys. Further, while calculating the attention scores, positional information is injected into the keys in the form of relative positional embedding as mentioned in [20].

$$a_{ab} = \text{softmax}_{x_{ab}} \left( \frac{q_{ij}^T \cdot k_{ab} + q_{ij}^T \cdot r_{a-i, b-j}}{\sqrt{h \times w}} \right) \quad (4)$$

Where  $r_{a-i,b-j}$  is obtained by concatenating row and column offset embeddings  $r_{a-i}$  and  $r_{b-j}$  respectively, with  $a-i$  and  $b-j$  being row and column offsets of each element  $ab \in \mathcal{N}_k$  from input  $x_{ij}$ . The attention weighted output  $y_{ij}^{att}$  corresponding to pixel  $x_{ij}$  is calculated as:

$$y_{ij}^{att} = \sum_{a,b \in \mathcal{N}_k} a_{ab} \cdot k_{ab} \quad (5)$$

Here, the same query and key matrices are used to calculate the attention outputs for each  $(i, j)$  of the input  $x$ .

Then we designed our model on the same structural backbone as our baseline CNN by replacing all the convolution layer with our self-attention layers.

**Attention as Activation Model:** Activation functions and attention mechanisms are typically treated as having different purposes and have evolved differently. However upon comparison, it can be seen that both the attention mechanism and the concept of activation functions give rise to non-linear adaptive gating functions [24]. To exploit both the locality of activation functions and the contextual aggregation of attention mechanisms, we use a local channel attention module, which aggregates point-wise cross-channel feature contextual information followed by sign-wise attention mechanism [24].

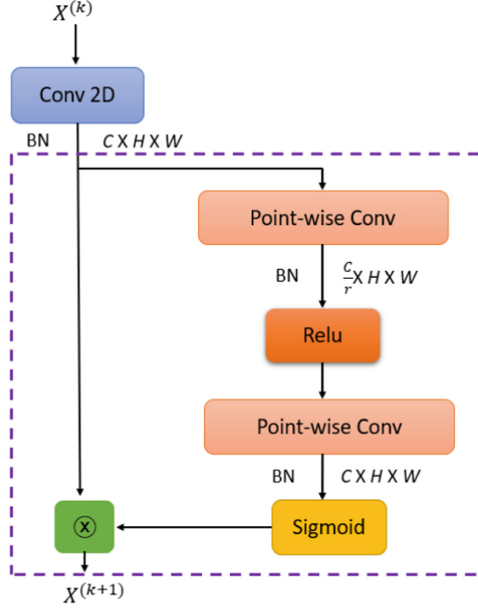
Our activation function resorts to point-wise convolutions [17] to realize local attention, which is a perfect fit since they map cross-channel correlations in a point-wise manner. The architecture of the local channel attention based attention activation unit is illustrated in Fig. 3. The goal is to enable the network to selectively and element-wisely activate and refine the features according to the point-wise cross-channel correlations. To reduce parameters, the attention weight  $L(X) \in R^{C \times H \times W}$  is computed via a bottleneck structure.

Input(X) is first passed through a 2-D convolutional layer into a point-wise convolution of kernel size  $\frac{C}{r} \times C \times 1 \times 1$  followed by batch normalization. The parameter  $r$  is the channel reduction ratio. This output is passed through a rectified linear unit (ReLU) activation function. The output of the ReLU is input to another point-wise convolution of kernel size  $C \times \frac{C}{r} \times 1 \times 1$  followed by batch normalization (BN in Fig. 3). Finally, to obtain the attention weight  $L(X)$ , the output is passed into a sigmoid function. It is to be noted that  $L(X)$  has the same shape as the input feature maps and can thus be used to activate and highlight the subtle details in a local manner, spatially and across channels. The activated feature map  $X'$  is obtained via an element-wise multiplication with  $L(X)$ :

$$X' = L(X) \otimes X \quad (6)$$

In element-wise sign-attention[23], positive and negative elements receive different amounts of attention. We can represent the output from the activation function ( $\mathcal{L}$ ) with parameters  $\alpha$  and  $X'$ .

$$\mathcal{L}(x_i, \alpha, X') = \begin{cases} C(\alpha)R(x_i), x_i < 0 \\ X'R(x_i), x_i \geq 0 \end{cases} \quad (7)$$



**Fig. 3.** Attention activation unit

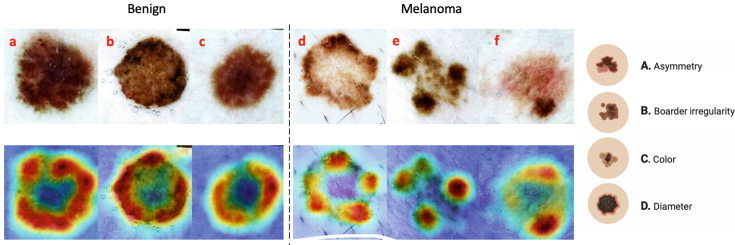
Where  $\alpha$  is a learnable parameter and  $C(\cdot)$  clamps the input variable between  $[0.01, 0.99]$ .  $X'$  is the above calculated activated feature map.  $R(X)$  is the output from standard rectified linear unit.

$$R(x_i) = \begin{cases} 0, & x_i < 0 \\ x_i, & x_i \geq 0 \end{cases} \quad (8)$$

This combination amplifies positive elements and suppresses negative ones. Thus, the activation function learns an element-wise residue for the activated elements with respect to ReLU which is an identity transformation, which helps mitigate gradient vanishing. We design the model based on our baseline CNN with only three blocks but with the above attentional activation function in place of ReLU.

### 2.3 ABCD Clinical Features and Classification

Dermatologists consider certain clinical features during the classification of malignant or benign skin lesions. A popular example is the **ABCDE** feature set [2]. In this approach, **A**symmetry, **B**order irregularity, **C**olor variation, **D**iameter and **E**volving or changing of a lesion region are taken into consideration for determining its malignancy (Ref. Fig. 4). **A**symmetry – Melanoma is often asymmetrical, which means the shape isn't uniform. Non-cancerous moles are typically uniform and symmetrical in shape. **B**order irregularity – Melanoma often has



**Fig. 4.** ABCD features used in dermatology diagnosis of skin lesions in dermatology.

borders that aren't well defined or are irregular in shape, whereas non-cancerous moles usually have smooth, well-defined borders. **Color variation** – Melanoma lesions are often more than one color or shade. Moles that are benign are typically one color. **Diameter** – Melanoma growths are normally larger than 6mm in diameter, which is about the diameter of a standard pencil. Since we do not have time series data, we extracted the first 4 (ABCD) features for each image in our dataset. Before feature extraction, an unsupervised segmentation framework is designed based on OTSU's [11] thresholding, morphological operations, and contour detection to separate out the main lesion region from the skin. From these segmented regions, the above-mentioned set of features were extracted using several transformations and elementary mathematical functions [12, 13]. Random Forest (RF) [14] and Support Vector Machines (SVM) [15] are used for the final classification with grid search [16] to find the optimal set of hyperparameters.

### 3 Experiments and Results

In this section, we present the experimental results, both quantitative and qualitative. First, in Table 1, we present the numerical results of the methods described in the preceding section. As evaluation metrics, we have used accuracy, AUC-ROC, precision, recall, and F1 score. Equalization sampling of minority classes was performed to tackle the problem of imbalanced dataset. All the deep learning models were trained to minimize the categorical crossentropy loss and the parameters were updated using ADAM optimizer [22].

First, we trained several traditional machine learning algorithms such as random forest and SVM, based on the ABCD features extracted as mentioned in Sect. 2.3. Grid search is used to choose the optimal set of hyperparameters and as shown in Table 1 a random forest model with 200 trees showed the best classification performance and its results are used for further comparison with the pixel-based models.

Next, multiple raw pixel-based deep learning models, as mentioned in Sect. 2.2, were trained and evaluated for the purpose of comparing and analyzing their performance with the ABCD feature based classification method, as well as to search for any feature correlation.



**Table 1.** Performance of different models with ABCD features and deep learned features

Method		Accuracy	AUC-ROC	Precision	Recall	F1 Score
Handcrafted feature based classification	Random forest on ABCD features	75.6	76.4	75	72	73
	SVM on ABCD features	74.4	75.6	74	71	72
Raw pixel feature based classification	Baseline CNN (from scratch )	78.3	69.4	72	67	69
	CNN with global attention	82.7	75.8	78	76	77
	Self-attention based model	74.2	68.1	71	67	69
	Attention as activation based model	71.4	68.8	68	66	67

**Table 2.** Performance measure of the variants of spatial self-attention layer

Variations of the spatial self-attention layer	Accuracy	AUC-ROC	Precision	Recall	F1 score
Original ( $k \neq v$ ) (Unscaled dot product) [20]	74.2	68.1	71	67	69
Proposed ( $k = v$ ) (Unscaled dot product)	74.5	67.4	71	69	70

### 3.1 Quantitative Results

Table 1 shows that even with suboptimal segmentation maps ABCD features have a high discriminating power of malignancy detection and classification. Further, use of finer lesion segmentation maps obtained by a manual or supervised approach can boost the classification performance of learning algorithms utilizing these sets of features. The overall performances of the deep models are also presented.

CNNs with global attention modules showed better results compared to the baseline CNN architecture that can be explained by the improved localization and feature selection capabilities of attention modules, whereas the self-attention based model performs similar to baseline CNN. Attention as activation based model outperforms CNNs of the same size. Self attention based models face the problem that using self-attention in the initial layers of a convolutional network yields worse results compared to using the convolution stem. This problem is overcome by Attention as activation based model and is the most cost effective solution as our activation units are responsible only for activating and refining the features extracted by convolution.

Table 2 shows that the performance of our proposed variation of the spatial self-attention model is not affected when we consider keys ( $k$ ) and values ( $v$ )

**Table 3.** Comparing alignment of deep models with ABCD features plus Random Forest

Method	Both correct (%)	ABCD features superior to raw pixel features (%)	Raw pixel features superior to ABCD features (%)	Both wrong (%)
Baseline CNN	70.2	5.4	8.1	16.3
CNN with attention	71.9	3.7	10.8	13.6
Self-attention based model	68.0	7.6	6.2	18.2
Attention as activation based model	66.5	8.9	5.5	19.1

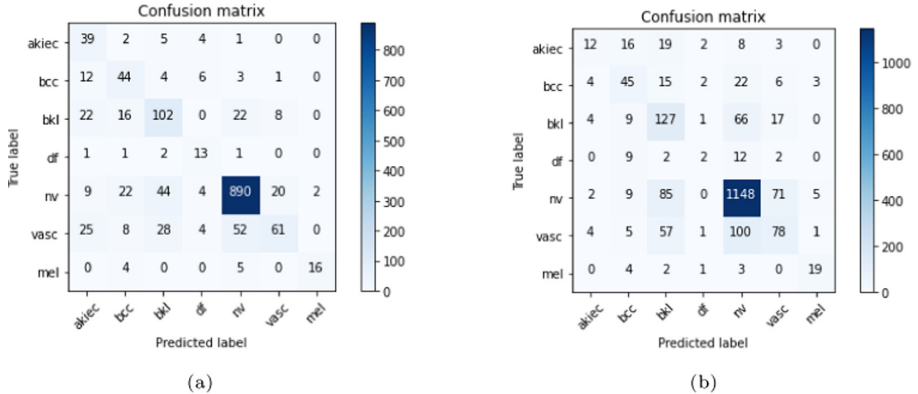
as identical metrics. This design of the spatial self-attention layer offers similar performance at lesser parameter settings and lower computational cost.

In Fig. 5, we present the confusion matrix of stand-alone self-attention and attention as activation models on the test dataset. Both models perform well on tumor types melanoma (mel), melanocytic nevi (nv), basal cell carcinoma (bcc), actinic keratosis (akiec), intraepithelial carcinoma and benign keratosis (bkl). However, occasionally the models confuse melanocytic nevi (nv) as benign keratosis (bkl) and vascular lesions (vasc) as melanocytic nevi lesions (nv).

### 3.2 Alignment Between ABCD Features and Deep Learned Features

To justify the decision level correlation between deep learned features and the ABCD features, the predictions on the test dataset were analyzed using four major criteria as presented in Table 3. We find relatively higher values in the first and last columns, where both the two broad classes of algorithms either succeed or fail, clearly indicating a correlation between their sought out features. Though this is not sufficient to establish direct feature correspondence, the results point towards some clinical relevance of deep models at a decision level.

We calculate the ABCD features from the attention maps of our self attention model and the ground truth segmentation maps. We use Random Forest and Support Vector Machine models on this data. The results are presented in Table 4. These results point towards the high correspondence in the ABCD features obtained by ground truth segmentation maps (clinical features) and the



**Fig. 5.** Confusion matrices for (a) Self-attention, and (b) Attention as activation. Both models confuse melanocytic nevi (nv) as benign keratosis (bkl) and vascular lesions (vasc) as melanocytic nevi lesions (nv).

attention maps of self attention based model (deep learned features). We also calculate the dice score [25] to compare the similarity between the ground truth segmentation maps and the deep learning model attention maps. The average dice score calculated over all the images as presented in Table 5. These positive results help us to closely examine how the raw pixel-based neural architectures associate with the clinical feature based learning algorithms at the feature level and indicate the similarity between model predicted and ground truth lesion regions. In a few failure cases, the dice score calculated was low. We present two such examples in Fig. 6.

**Table 4.** Performance measure of ABCD features learned from ground truth segmentation maps and self-attention based model

Method	Accuracy	F1 score	Recall	Precision
Ground truth segmentation maps	70	60	70	62
Attention map of self-attention based model	67	54	67	45

### 3.3 Qualitative Results

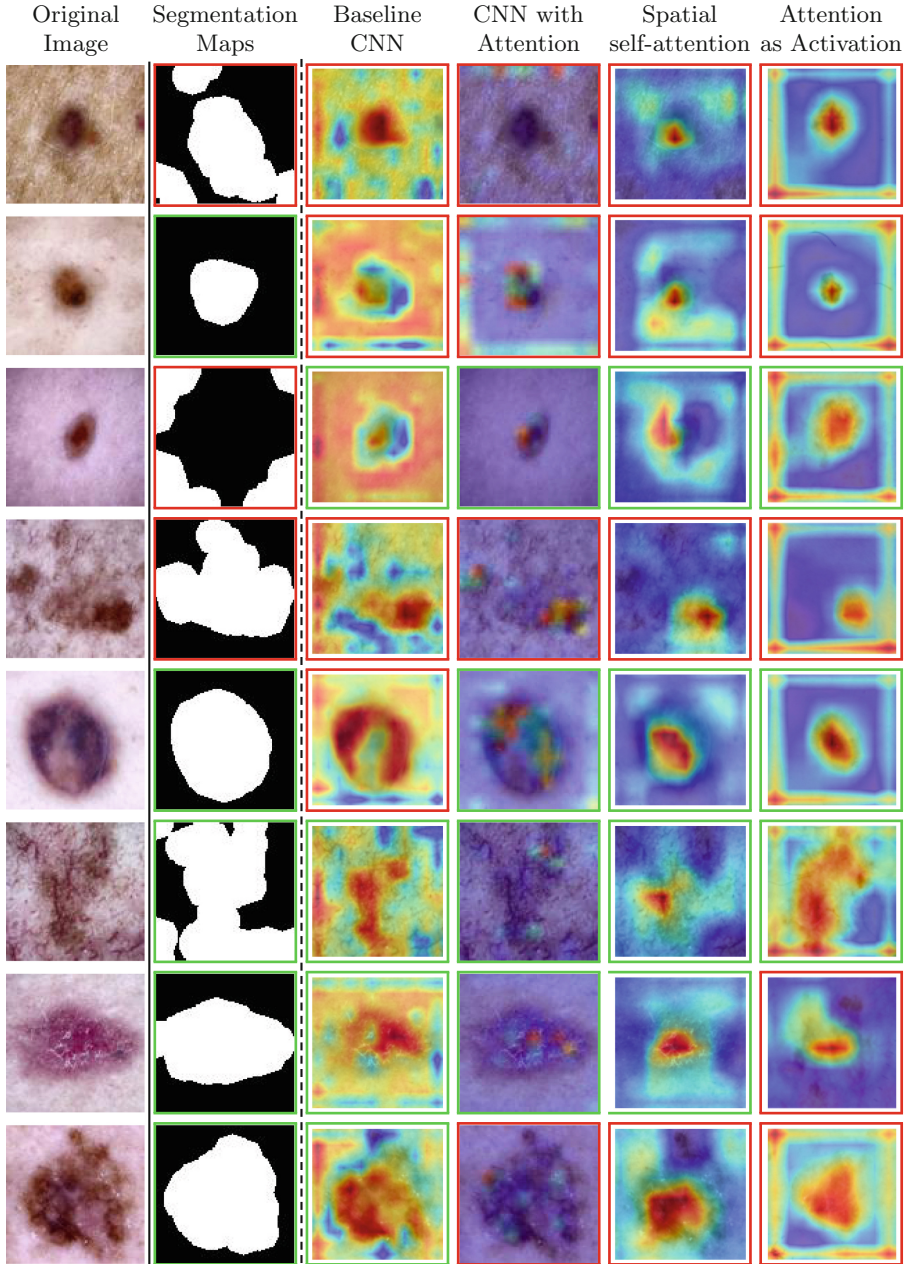
Next, we have visually explored whether there is any direct alignment between the deep learned features and ABCD features by analyzing their global feature descriptors and segmentation maps, respectively, for a random set of test images.

**Table 5.** Dice score between the ground truth segmentation maps and deep architectures

Method	Dice score
Baseline CNN	79.5
Self-attention based model	73.9
Attention as activation based model	74.6

**Fig. 6.** (a) shows an example of correct segmentation with high dice score of 0.96. (b) is an example of incorrect output with low dice score 0.01. For each pair, we present the ground truth on the left and the model output on the right.

CAM [21] is used for visualizing the global feature descriptors for the deep classification models. From the visual results presented in Fig. 7, it is clear that the ability to precisely localize the lesion region is the most crucial quality that a model should possess. For most of the cases, whenever the attention heat maps have a satisfactory overlapping with the correct segmentation map (rows 3, 5, 6, 7) the results are correct, and whenever they differ significantly (row 2) the results are incorrect. The third column of the figure shows the activation maps of the baseline CNN to be very sparse that indicates poor localization capability, leading to many incorrect predictions. The localization capability of the attention-based models (columns 4,5 and 6) are much better than the baseline CNN that accounts for better classification results. These attention-based models have helped to pinpoint the lesion areas in the image and better addressed the fine-grain nature of the problem. Visually the localization power of the spatial self-attention and attention as activation models are quite accurate, however, in many cases, they tend to focus on the boundary regions of the image or have poor overlapping with the lesion area, which leads to incorrect predictions and suboptimal results. A good dice score suggests a descent alignment of model activations with some of the clinical features such as Asymmetry and Border irregularity, reflecting with their accuracy.



**Fig. 7.** Comparison of segmentation maps used for ABCD feature extraction and important regions according to deep learning models for a random set of test images. A red box around a segmentation/attention map represents incorrect prediction whereas a green box denotes correct prediction. (Color figure online)

## 4 Conclusion

In this work, we have investigated whether the features extracted by deep models such as convolutional networks, self-attention models and attention as activation models correlate with clinically relevant features. We have taken automated skin cancer detection as the test case and the quantitative, as well as qualitative results, point towards an underlying correlation between them at feature and decision level. A visual analysis has been performed to check whether the activation maps of deep models do possess any similarity with the segmentation maps used for clinical feature (ABCD features for skin lesion) extraction. Where the clinical features are unique and concrete representations of a lesion region, the deep learned features are more abstract and compound. However, with the help of a comparative analysis of different methods we are able to bridge the gap of trustability, when it comes to justifying their output.

## References

1. LeCun, Y., Bengio, Y.: Convolutional networks for images, speech, and time series. In: *The Handbook of Brain Theory and Neural Networks* (1995)
2. Jensen, D., Elewski, B.E.: The ABCDEF rule: combining the ABCDE rule and the ugly duckling sign in an effort to improve patient self-screening examinations. *J. Clin. Aesthetic Dermatol.* **8**(2), 15 (2015)
3. Van Molle, P., De Strooper, M., Verbelen, T., Vankeirsbilck, B., Simoens, P., Dhoedt, B.: Visualizing convolutional neural networks to improve decision support for skin lesion classification. In: Stoyanov, D., et al. (eds.) *MLCN/DLF/IMIMIC -2018. LNCS*, vol. 11038, pp. 115–123. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-02628-8\\_13](https://doi.org/10.1007/978-3-030-02628-8_13)
4. Young, K., Booth, G., Simpson, B., Dutton, R., Shrapnel, S.: Deep neural network or dermatologist? In: Suzuki, K., et al. (eds.) *ML-CDS/IMIMIC -2019. LNCS*, vol. 11797, pp. 48–55. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-33850-3\\_6](https://doi.org/10.1007/978-3-030-33850-3_6)
5. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626 (2017)
6. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*, pp. 4765–4774 (2017)
7. Aggarwal, A., Das, N., Sreedevi, I.: Attention-guided deep convolutional neural networks for skin cancer classification. In: *IEEE International Conference on Image Processing Theory, Tools and Applications*, pp. 1–6 (2019)
8. Zhang, J., Xie, Y., Xia, Y., Shen, C.: Attention residual learning for skin lesion classification. *IEEE Trans. Med. Imaging* **38**(9), 2092–2103 (2019)
9. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **5**(1), 1–9 (2018)
10. Pizer, S.M., et al.: Adaptive histogram equalization and its variations. *Comput. Vis. Graphics Image Process.* **39**(3), 355–368 (1987)

11. Otsu, N.: A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**(1), 62–66 (1979)
12. Zaqout, I.: Diagnosis of skin lesions based on dermoscopic images using image processing techniques. *Pattern Recognition-Selected Methods and Applications Intech Open* (2019)
13. Amaliah, B., Faticah, C., Widyanto, M.R.: ABCD feature extraction of image dermatoscopic based on morphology analysis for melanoma skin cancer diagnosis. *Jurnal Ilmu Komputer dan Informasi* **3**(2), 82–90 (2010)
14. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
15. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**(3), 273–297 (1995)
16. Bergstra, J., Bardenet, R., Bengio, Y., Kégl, B.: Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems* (2011)
17. Lin, M., Chen, Q., Yan, S.: Network in network. [arXiv:1312.4400](https://arxiv.org/abs/1312.4400) (2013)
18. Jetley, S., Lord, N.A., Lee, N., Torr, P.H.S.: Learn to pay attention. [arXiv:1804.02391](https://arxiv.org/abs/1804.02391) (2018)
19. Vaswani, A., et al.: Attention is all you need. In: *Advances in Neural Information Processing Systems* (2017)
20. Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-alone self-attention in vision models. [arXiv:1906.05909](https://arxiv.org/abs/1906.05909) (2019)
21. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016)
22. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: *International Conference on Learning Representations* (2014)
23. Chen, D., Li, J., Xu, K.: AReLU: attention-based rectified linear unit. [arXiv:2006.13858](https://arxiv.org/abs/2006.13858) (2020)
24. Dai, Y., Oehmcke, S., Gieseke, F., Wu, Y., Barnard, K.: Attention as activation. [arXiv:2007.07729](https://arxiv.org/abs/2007.07729) (2020)
25. Eelbode, T., et al.: Optimization for medical image segmentation: theory and practice when evaluating with Dice score or Jaccard index. *IEEE Trans. Med. Imaging* **39**(11), 3679–3690 (2020)
26. Nida, N., Irtaza, A., Javed, A., Yousaf, M.H., Mahmood, M.T.: Melanoma lesion detection and segmentation using deep region based convolutional neural network and fuzzy C-means clustering. *Int. J. Med. Inform.* **124**, 37–48 (2019)
27. Bisla, D., Choromanska, A., Berman, R.S., Stein, J.A., Polsky, D.: Towards automated melanoma detection with deep learning: data purification and augmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2019)
28. Adekanmi, A.A., Viriri, S.: Deep learning-based system for automatic melanoma detection. *IEEE Access* **8**, 7160–7172 (2019)
29. Adekanmi, A.A., Viriri, S.: Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. *Artif. Intell. Rev.* **54**(2), 811–841 (2021)
30. Codella, N., et al.: Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC). [arXiv:1902.03368](https://arxiv.org/abs/1902.03368) (2019)