On Interpretability of Deep Learning based Skin Lesion Classifiers using Concept Activation Vectors

Adriano Lucieri
Smart Data and Knowledge Services
German Research Center for Artificial Intelligence GmbH (DFKI)
Kaiserslautern, Germany
adriano.lucieri@dfki.de

Muhammad Naseer Bajwa
Technische Universität Kaiserslautern
German Research Center for Artificial Intelligence GmbH (DFKI)
Kaiserslautern, Germany
0000-0002-4821-1056

Stephan Alexander Braun
Universitätsklinikum Münster
Universitätsklinikum Düsseldorf
Germany
0000-0002-5600-0195

Muhammad Imran Malik
School of Electrical Engineering and Computer Science
National University of Science and Technology
Islamabad, Pakistan
0000-0002-8079-5119

Andreas Dengel
Technische Universität Kaiserslautern
German Research Center for Artificial Intelligence GmbH (DFKI)
Kaiserslautern, Germany
0000-0002-6100-8255

Sheraz Ahmed
Smart Data and Knowledge Services
German Research Center for Artificial Intelligence GmbH (DFKI)
Kaiserslautern, Germany
0000-0002-4239-6520

***Abstract*—Deep learning based medical image classifiers have shown remarkable prowess in various application areas like ophthalmology, dermatology, pathology, and radiology. However, the acceptance of these Computer-Aided Diagnosis (CAD) systems in real clinical setups is severely limited primarily because their decision-making process remains largely obscure. This work aims at elucidating a deep learning based medical image classifier by verifying that the model learns and utilizes similar disease-related concepts as described and employed by human dermatologists. We used a well-trained and high performing neural network developed by REasoning for COmplex Data (RECOD) Lab for classification of three skin tumours, i.e. Melanocytic Naevi, Melanoma and Seborrheic Keratosis and performed a detailed analysis on its latent space. Two well established and publicly available skin disease datasets, PH^2 and derm7pt, are used for experimentation. These datasets provide annotations for various disease-related concepts in addition to image-wise diagnostic labels. Human understandable concepts are mapped to RECOD image classification model with the help of Concept Activation Vectors (CAV). Our results on independent evaluation sets clearly show that the classifier learns and encodes human understandable concepts in its latent representation. Additionally, TCAV scores (Testing with CAV) suggest that the neural network indeed makes use of disease-related concepts in the correct way when making predictions. We anticipate that this work can not only increase confidence of medical practitioners on CAD but also serve as a stepping stone for further development of CAV-based neural network interpretation methods.**

***Index Terms*—Skin Lesion Classification, Medical Imaging, Computer-Aided Diagnosis, Explainable Artificial Intelligence, Concept Activation Vectors, Convolutional Neural Networks.**

I. Introduction

United Nations (UN) has recognised healthcare and well-being as one of the 17 Sustainable Development Goals (SDGs) to create a better future for all by 2030 [1]. However, achieving this goal requires concerted and sustained efforts in utilizing all available means to improve healthcare since many people are needlessly suffering from preventable diseases. In 2017, AI for Good, a UN initiative to provide a global platform for researchers, identified great potential of Artificial Intelligence (AI) to achieve these SDGs and to help solve the greatest global challenges [2].

Numerous remarkable studies have been conducted in the last few years successfully applying deep learning for disease classification using various medical image modalities [3]–[5]. However, the acceptance of such Computer-Aided Diagnosis (CAD) solutions with doctors and patients remains dubious at best due to the fact that the process behind learning and encoding features in latent space by computer models is not well understood. This lack of transparency in the whole decision-making process cannot be overlooked in various critical application areas including medical diagnosis. Especially after Europe's General Data Protection Regulations (GDPR) [6] came into effect in 2018, data subjects are entitled to *Right to Explanation* for any automated decision made by computer algorithms. It is, therefore, need of the hour to elucidate the working principle of deep learning based classifiers so that practical applications of AI in medical diagnosis can be realized expeditiously.

Compared to other fields of applications of Deep Neural Networks (DNNs), medical image analysis often presents unique challenges due to inherent complexity of this task. Manual classification of complex diseases involves recognizing subtle features and high-level concepts that are challenging to grasp without expert knowledge. Even with expert knowledge, doctors' subjective understanding of biomarkers leads to low inter-expert agreement [7], [8]. Therefore, common

This work is partially funded by National University of Science and Technology (NUST), Pakistan through Prime Minister's Programme for Development of PhDs in Science and Technology.

explanation methods like visualization of saliency maps, that strongly rely on spatial divisibility of concepts, work well on common object detection tasks [9]–[11] that have well distinguishable features, but fail on more complex medical image analysis tasks.

Skin cancer is the most common type of cancer in the U.S [12]. According to a recent study [13], skin cancer related death rate forecast for U.S. in 2019 amounts to 11,650 people. These rising rates of skin cancer incidences can not only cost precious lives but also incur huge burden on healthcare system. It estimated that approximately 3 million people are treated annually for skin cancer in the U.S. and it costs around 8.1 billion USD [14].

In this study, we chose classification of skin diseases as a use case to understand what DNNs learn and what they rely on for their predictions in medical diagnosis. We attempt to understand if the *concepts* learnt by classifiers in complex medical imaging tasks are similar to those used by dermatologists. We used two publicly available datasets of dermoscopic images i.e. PH^2 and derm7pt. These datasets were selected because they provide concept annotations in addition to image-wise diagnosis labels. Summarizing our contributions, this study presents

- Mapping of concepts learnt by a deep model, in its latent space, for skin lesion classification to dermatologically significant human-understandable concepts using Concept Activation Vectors (CAVs).
- Analysis of different dermoscopic criteria's contribution to the deep model's predictions, revealing agreement between doctors' reasoning process and that of deep model.
- Analysis of potential bias introduced by training CAVs using identically distributed dataset samples.
- Critical analysis of the TCAV score from TCAV (Testing with CAVs) interpretability method [15].

II. Related Work

Understanding the way neural networks learn and explaining their prediction behaviour are active areas of research [16], [17]. To interpret these inherently nonlinear mathematical models, three main types of methods are usually employed.

A. Saliency-Based Neural Network Explanations

Saliency-based methods for neural network explanation were among the first tools towards explainable AI and are still highly in use today. Examples for those methods are GradCAM [11], SmoothGrad [18], Integrated Gradient [19] and Layer-Wise Relevance Propagation (LRP) [20] to name a few. Since these methods create activation maps based on individual input samples, they provide only local interpretations and are unable to explain network's decisions on a global dataset scale.

To date, there are very few works focusing on interpreting deep classifiers for dermoscopic images using saliency-based attribution methods [21], [22]. This might partly be due to the innate difficulty in skin lesion classification that mandates a huge amount of expert knowledge to recognize complex and subtle structures. It could also be due to a large variation in fine nuances of these structures that are hard to discern yet can drastically change diagnosis. Moreover, the visual artefacts corresponding to various diseases in skin lesion images sometimes overlap and usually are distributed all over the image, which does not fare well with saliency-based model interpretation. Other domains in which these methods are frequently used for interpretation like object classification [9]–[11] usually show more discriminatory features corresponding to specific parts of an object – for instance tires or headlights of cars.

B. Text-Based Neural Network Explanations

Textual explanation methods for neural networks can be either template-based [23]–[25], that generate justifications from some auxiliary information or rule-based [26]–[29], where a classifier is trained with images as well as additional natural language explanations. Zhang et al. [28] proposed a unified network called MDNet following rule-based approach that generates diagnostic reports along with corresponding attention maps of input images in order to increase the semantic and visual interpretability of the medical imaging task at hand. Jing et al. [29] proposed a multi-task learning framework that is able to localize abnormal regions in medical images, predict tags and generate descriptions thereof.

C. Concept-Based Neural Network Explanations

Concept-based explanation addresses the problem of explaining black-box models by mapping human-understandable concepts to neural networks' latent representation. Kim et al. [15] introduced *Concept Activation Vectors* (CAVs) which are calculated in a supervised way using general human concept patches that were taken from various domains. By calculating these main concept directions and by leveraging directional derivatives, they were able to quantify the influence of a concept to specific output classes. Zhou et al. [30] followed a similar approach by decomposing neural networks' activations into semantically meaningful components pre-trained from a large concept corpus. Graziani et al. [31], [32] restated the problem of classification in CAVs and employed regression instead. They applied their so-called *Regression Concept Vectors* (RCVs) on problems from medical domain like binary classification of breast cancer histopathology slides and classification of Retinopathy of Prematurity (ROP) states. As an extension to the work in [15], Ghorbani et al. [33] recently developed a method for an unsupervised clustering of object datasets by first applying segmentation of single objects and then clustering the object patches' activations into semantically meaningful clusters.

To the best of our knowledge, concept-based explanation methods have not previously been explored for skin lesion classification networks. Due to the nature of this problem, not all of the previously described methods can be directly applied on this task. Unsupervised clustering as used in [33], for example, is not suitable in skin lesions as there is a huge spatial concept overlap and thus no possibility for distinct part segmentation. RCVs are also not applicable as skin lesion concepts are hardly quantifiable. The method in [30] requires a concept corpus which is not readily available for this specific task. Any type of textual explanation generation is also not applicable, as no diagnostic reports or descriptions of diagnosis are provided with any dermoscopic skin lesion dataset. The computation of CAVs as seen in [15] requires patches corresponding to general human-understandable concepts. In this work we adopt the TCAV method to the problem of skin lesion classification. Instead of providing general, out-of-distribution concept patches, we train CAVs using samples from identically distributed datasets to map human-understandable concepts to the network's latent space. Moreover, we study the applicability of this adapted method to skin lesion classification.

III. Background

This section briefly describes CAVs and the method of calculating TCAV scores which are used in this work to quantify the contribution of a concept to DNN's prediction.

A. Concept Activation Vectors

In order to achieve human-centered interpretability of DNNs, Kim et al. [15] introduced *Concept Activation Vectors*. A CAV, $\vec{v}_c$, is a vector in the embedding space of a neural network pointing into the direction that encodes the given concept c. CAVs can be calculated by training a binary concept classifier dividing samples containing a given concept from samples where the concept is absent. The CAV is then defined as the normal of the hyperplane separating the two classes.

a) TCAV Score: The metric introduced in [15] to score the influence of a CAV on a class of input images is the TCAV score. It makes use of directional derivatives $S_{C,k,l}(x)$ to measure the contextual sensitivity of a concept towards an entire input class, therefore providing global explanations. The TCAV score is given by:

$$TCAV_{Q_{C,k,l}} = \frac{|x \in X_k : S_{C,k,l}(x) > 0|}{|X_k|} \tag{1}$$

where X_k denotes all inputs, k represents the class labels and $S_{C,k,l}(x)$ the directional derivate of a sample's activation x from layer l with respect to class k and concept C. The TCAV score effectively measures the ratio of class k's inputs, that are positively affected by concept C without taking any magnitude into account. As compared to saliency maps or other per-feature metrics, the TCAV score allows for quantitative evaluation of concepts on whole input classes.

B. Dermoscopic Concepts used for Analysis

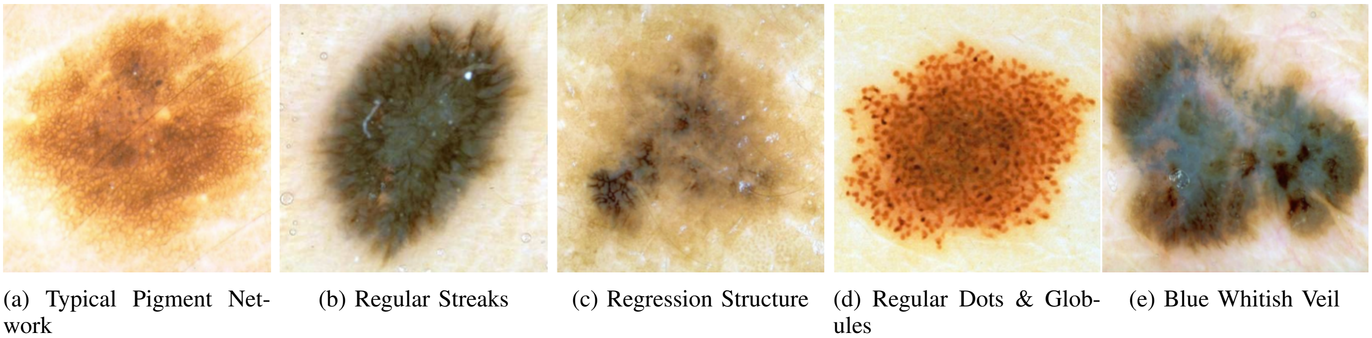
The concepts used in this work to interpret the deep classifier are briefly defined below in accordance with standardized terminology agreed upon by expert dermatologist in 3rd Consensus Conference of the International Society of Dermoscopy (IDS) [34]. Fig. 1 depicts examples of some concepts listed below.

1) Pigment Networks: Pigment Networks consist of interconnected pigmented lines forming a gridlike pattern. Depending on the subtype of Pigment Network, it can either have minimal variability in colour, thickness and spacing of the lines, forming a symmetric grid (Typical Pigment Network) or have greater variability in colour, thickness and spacing of the lines, forming an asymmetric grid (Atypical Pigment Network). Apart from those two general types, more subtypes are also defined in literature. Atypical Pigment Networks can be a clue for Melanoma (although many dysplastic naevi also have atypical networks) whereas typical Pigment Networks normally indicate benign melanocytic lesions (Naevi).

2) Streaks: Streaks describe an abnormality of the lesion that can either have the form of pure straight radial extensions, radial extensions with bulbous and often kinked projections on their ends or a widening of broken lines with incomplete connections. Streaks are referred to as irregular if they are irregularly distributed along the edge of the lesion and are brown-black in colour [35]. Regular Streaks indicate benign lesions and Irregular Streaks are clues for malignant Melanoma.

3) Regression Structures: Regression Structures are characterized by the appearance of either areas of fine, grey-blue dots, or areas of skin whiter than the surrounding normal-looking skin without blood vessels or shiny-white structures. Its presence is highly indicative of melanoma [35].

4) Dots and Globules: Dots are small structures of pigmented areas clustered in any distribution on or around the lesion. Dots clustered in center regions or on the network



(a) Typical Pigment Network

(b) Regular Streaks

(c) Regression Structure

(d) Regular Dots & Globules

(e) Blue Whitish Veil

Fig. 1: Some exemplary cases of skin lesion concepts from derm7pt dataset

lines are referred to as regular, otherwise irregular. Globules are round, oval or polygonal structures larger than dots that can have high variability in colour, size and shape along with asymmetric distribution (Irregular Globules) or minimal variability along with symmetric distribution (Regular Globules). Dots and Globules are indicators for benign melanocytic lesions if they are regular and for melanoma if they are irregular [35].

5) Blue-Whitish Veils: Blue-Whitish Veils describe an irregularly shaped, structureless blotch on the lesion area that is characterized by a blue hue with an overlying whitish ground-glass haze. In [36] it is rated as the most useful single diagnostic indicator for melanoma.

6) Asymmetry: Asymmetry is the most important factor in malignant melanoma identification using ABCD rule [37]. In this work, asymmetry refers to an asymmetrical lesion contour as well as asymmetrical distributions of structures and colours within a lesion [38]. The asymmetry concept is further divided into asymmetry in one or two axes.

7) Colour: This concept refers to colour present within the lesion area. As the appearance of single colours is not yet indicative of any diagnosis, a combined concept of *three or more colours* is used in the analysis. The presence of three or more colours increases the probability of melanoma drastically [35].

The intricate explanations of concepts given above along with the concepts' innate variability offer much room for interpretation, implying the complexity of the problem itself. This is evident by the fact that even doctors tend to have notable disagreements when it comes to diagnosis, localization or identification of concept [7], [8].

IV. Materials & Method

A. Model

The model used in this work as the basis for our exploration and experimentation is developed by the University of Campinas in Brazil. Their RECOD Lab (REasoning for COmplex Data) made their submission [39] to the IEEE International Symposium on Biomedical Imaging (ISBI) 2017 challenge publicly available on github[1]. By applying a transfer learning approach combined with extensive ensembling using an SVM meta-layer on top of seven base models trained on different data subsets, they achieved best Area Under Receiver Operating Characteristic (ROC) Curve (AUC) for Melanoma (MEL) classification (87.4%), 3rd best AUC for Seborrheic Keratosis (SK) classification (94.3%), and 3rd best combined/mean AUC (90.8%) in part 3 of 2017 challenge. In this work, we intentionally refrained from training our own skin lesion classification model as our primary objective was explainability of these models instead of their classification performance. Thus, for our experimentation we only focused on a single module from RECOD's well-trained architecture. We used one of the base models[2] with Inception v4 [40] architecture, subsequently referred to as *model* or *network*. It was trained on RECOD's "deploy" set of 9,640 images using per-image normalization. For all experiments, network's activations from the second last convolutional layer *mixed_6h* are used for concept mapping.

TABLE I: Distribution of image samples into different concept classes in PH^2 and derm7pt datasets. Note that PH^2 dataset does not distinguish between regular and irregular streaks.

Concepts	**Presentation**		**Abbreviation**	**PH^2 [38]**	**derm7pt [41]**
Pigment Network	Absent		PN_A	0	272
	Present	Typical	PN_T	84	335
		Atypical	PN_AT	116	216
Streaks	Absent		ST_A	170	490
	Present	Regular	ST_R	30	96
		Irregular	ST_IR		237
Regression Structures	Absent		RS_A	175	590
	Present		RS	25	233
Dots & Globules	Absent		DG_A	87	133
	Present	Regular	DG_R	54	300
		Irregular	DR_IR	59	390
Blue-Whitish Veils	Absent		BWV_A	36	182
	Present		BWV	164	641
Asymmetry	Absent		Asym_A	117	N/A
	Present	1-Axis	Asym_1	31	N/A
		2-Axis	Asym_2	52	N/A
Colours	2 or fewer		C_2-	161	N/A
	3 or more		C_3	39	N/A
Total Samples				200	823

B. Datasets

The datasets used for concept training are PH^2 dataset [38] and Seven-Point Checklist Dermatology dataset abbreviated as derm7pt [41].

The PH^2 dataset consists of 200 dermoscopic images of melanocytic lesions, including 80 common naevi, 80 atypical naevi, and 40 melanomas. Along with the images, colour and lesion segmentation masks are provided as well as extensive well-curated annotations as seen in Table I. The derm7pt dataset consists of 1,011 clinical and dermoscopic images. Each sample is assigned to either a miscellaneous class or one of 4 diagnosis classes. Two of these diagnosis classes i.e. Melanoma and Naevi (NV) are further divided into 13 subclasses. From this dataset, only MEL and NV samples have been considered, resulting in 823 images. SK samples have been discounted due to their low count of only 45 samples. Table I provides an overview of number of samples for each concept class.

For evaluation purposes, the original ISBI 2017 challenge test and train sets [42] are used. The train set of ISBI 2017 challenge contains 1372 samples of NV, 374 samples of MEL and 254 samples of SK whereas test set contains 393 images of NV, 117 images of MEL and 90 images of SK.

[1] https://github.com/learningtitans/isbi2017-part3

[2] *checkpoint.rc25*

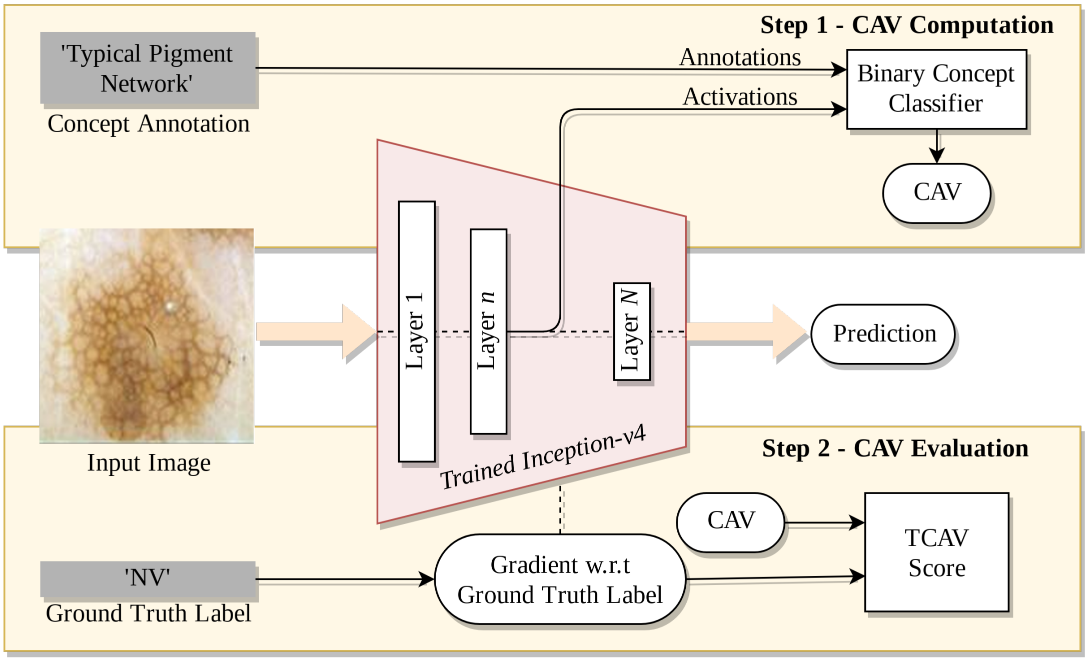


Fig. 2: Overview of training concept classifiers and calculating CAV and TCAV scores

In order to verify statistical significance of our results, we trained random CAVs to compare against our concept CAVs. For this purpose, a subset of the ISIC archive[3] images is used, excluding MEL and NV classes, resulting in 2,870 samples. The idea behind leaving out those two classes is that remaining samples hardly contain concepts similar to the ones used for concept training.

C. Experimental Setup

As previously described, all experiments have been conducted on one of the Inception v4 base models from [39]. For each concept, binary classifiers are trained on network's activations to find the concepts' directions in the embedding space. The training and evaluation scheme is depicted in Fig. 2. First, activations are extracted from seven locations of the Inception v4 model using PH^2 and derm7pt datasets. A clustering-based under-sampling technique as well as stratified splitting is applied to ensure evenly balanced train and validation sets for each binary concept training. Second, TCAV score is used to evaluate a concept's importance to a specific target class. To account for differences in pre-processing and classifier initialization, each classifier training is repeated 100 times on a randomly sampled dataset split, resulting in different CAVs and different TCAV scores.

To check statistical significance of learned concepts, we trained additional 100 random CAVs per layer. The random datasets are produced by repeatedly sampling 1,000 random images from ISIC archive subset described in section IV-B and assigning random binary labels. The distribution of random concept TCAV scores and actual concept TCAV scores is then compared by conducting a two-sided t-test with $\alpha = 0.05$ to assure significance of the found CAVs. In the results section, statistical insignificance is represented by red asterisks on top of the plotted bars.

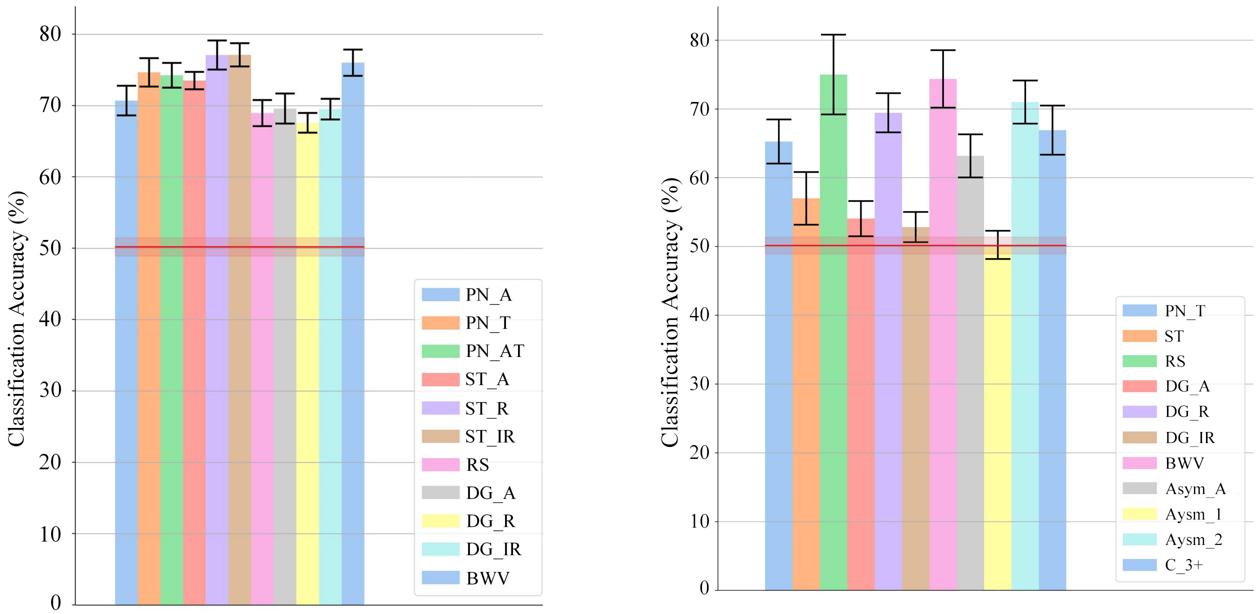
V. Results and Analysis

Fig. 3a shows all mean validation accuracies achieved by individual binary concept classifiers along with their standard deviation trained on derm7pt embeddings from last convolutional layer. Mean baseline results from training on 50 random concept subsets are depicted by horizontal red line along with light red surrounding area marking variance. It is evident from the figure that all concept classifiers achieved significantly higher validation accuracies than random baseline. The overall accuracies achieved might not seem very high, however it has to be mentioned here that computation of CAV requires that only linear classifiers may be used since it's easier to calculate normal vector if the decision surface is a hyperplane. Using non-linear classifier can improve accuracies but it will hinder CAV calculation. Therefore, keeping in view linear concept classifiers, the results are clear evidence that network's latent space is structured in a way that allows activation's separation with respect to concepts.

Fig. 3b shows the classifiers' validation accuracy trained on PH^2 dataset concepts on the network's last layer. It is notable that many concepts achieved relatively mediocre accuracies near the random baseline. This can be explained by very small number of positive concept samples available in this dataset.

The TCAV score quantifies positive or negative influence of a given concept towards a specific target class. Values above 0.5 indicate positive influence of the concept to the prediction and lower values indicate negative influence. Figures 4 and 5

[3] https://isic-archive.com/



(a) Derm7pt dataset

(b) PH^2 dataset

Fig. 3: Validation accuracies of all concept classifiers trained and tested individually on derm7pt and PH^2. Random baseline is denoted by horizontal red line along with light red area marking standard deviation. Insignificant classifiers are marked with a red asterisk.

show the TCAV scores achieved by evaluating 20 CAVs per concept on the last layer using both datasets. Average baseline scores of all 50 random concepts are again depicted by red horizontal lines along with their standard deviation in light red. Statistically insignificant results are marked by red asterisks.

Results for NV and MEL classes for concepts trained using derm7pt look very much as expected. Although, scores for *PN_A* and *DG_A* turned out to be insignificant in the

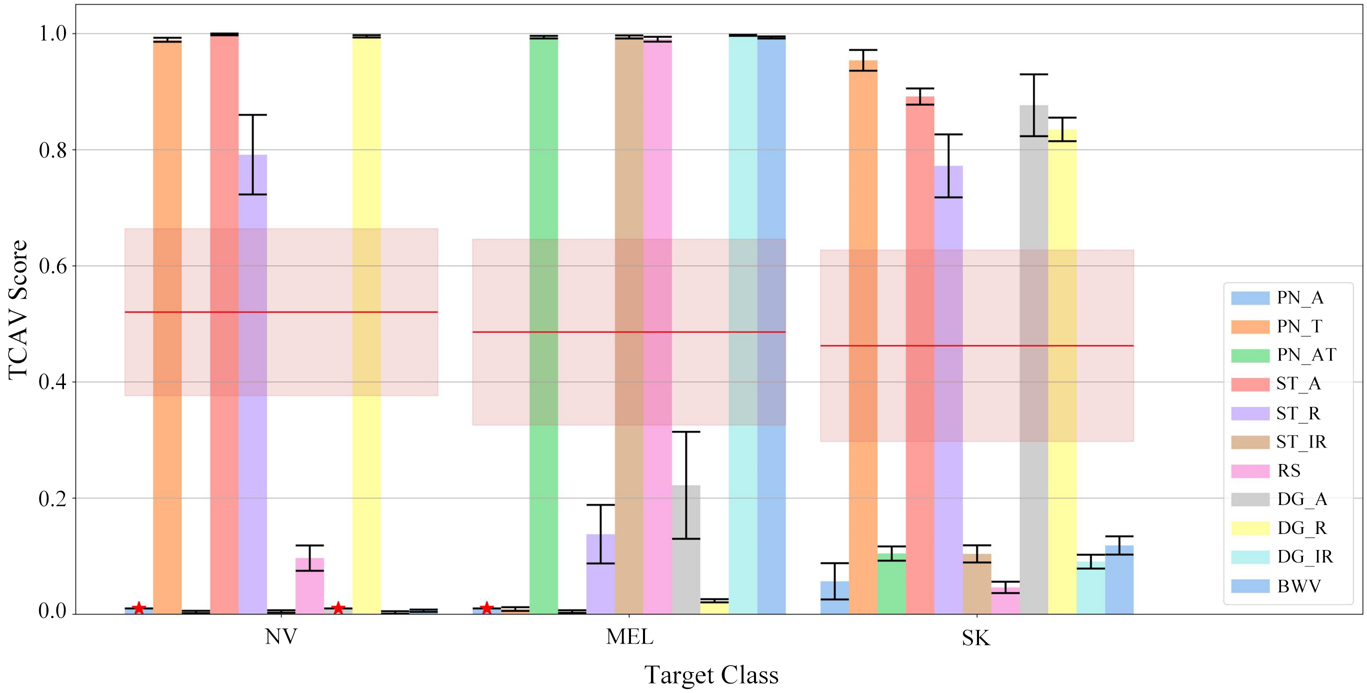


Fig. 4: TCAV scores of each concept for derm7pt with respect to each target class on the last layer of RECOD model.

experiments, features indicating benign melanocytic lesions like *PN_T*, *ST_A*, *ST_R* and *DG_R* all contributed positively towards NV class. Also it is notable that the total absence of Streaks (*ST_A*) has a stronger positive influence as compared to the presence of regular Streaks (*ST_R*). On the other hand, strong signs for malignant melanoma like *PN_AT*, *ST_IR*, *RS*, *DG_IR* and *BWV* show strong negative influence. Results for MEL class show the exact opposite behaviour, which is perfectly aligned with the descriptions in medical literature. It is again notable that the absence of Dots and Globules (*DG_A*) and the presence of regular Streaks (*ST_R*) show less negative impact to MEL class as compared to, for example, absence of Streaks (*ST_A*). Results for SK class show positive influence for typical Pigment Networks (*PN_T*). In [43] the appearance of network structures that resemble those of benign melanocytic lesions in Seborrheic Keratosis has been confirmed. In the same study, evidence for Dots in SK lesions has also been found.

Fig. 5 shows resulting TCAV scores for CAVs trained on PH^2 dataset. All concepts achieving less than 55% validation accuracy have not been considered. Again, TCAV scores for NV and MEL show expected behaviour. Only typical Pigment Networks (*PN_T*), typical Dots and Globules (*DG_T*) and full Symmetry (*Asym_A*) contribute positively towards Naevi class. Additionally, from the results it appears that lesions containing fewer than three colours also tend to be classified as benign melanocytic naevi. For melanoma, the exact opposite holds again which can be confirmed by the concept descriptions in Section III-B. For SK we can again observe high influence of typical Pigment Networks as well as full Symmetry.

To further validate that the model has comprehensively learnt these disease-related concepts instead of learning some random concepts, we had our model sort all the test images with respect to degree of visibility of a certain concept in each image. In other words, the model ordered all 300 test images starting from those that presented very obvious existence of a concept and ending with those which had least evidence of that concept. This ordering in performed based on euclidean distance in a CAV's direction. Fig 6 through Fig. 8 show first five and last five images from the sorted test set with respect to different concepts. The first row of each figure shows positive examples, where the concept is most clearly visible, and the second row shows negative examples, where the concept is virtually absent. It is evident from these figures that the proposed method for explaining skin disease classifiers does not only provide justification of classifier's decision on global dataset scale but also sensibly identified reasons for per-image predictions.

VI. Conclusion

Concept-based methods for network explanation offer great potential especially for complex classification tasks in critical application areas like medical image analysis. This work strives to leverage these methods to verify deep neural networks' ability to learn and utilize human understandable concepts for skin lesion classification. Our results show strong correlation between DNN's learnt representation of various concepts and those routinely used by human dermatologists. This work corroborates that deep learning based CAD systems are able to learn and utilize same disease-related concepts for prediction as used by dermatologists. We hope that physicians would be able to confer higher confidence to such CAD

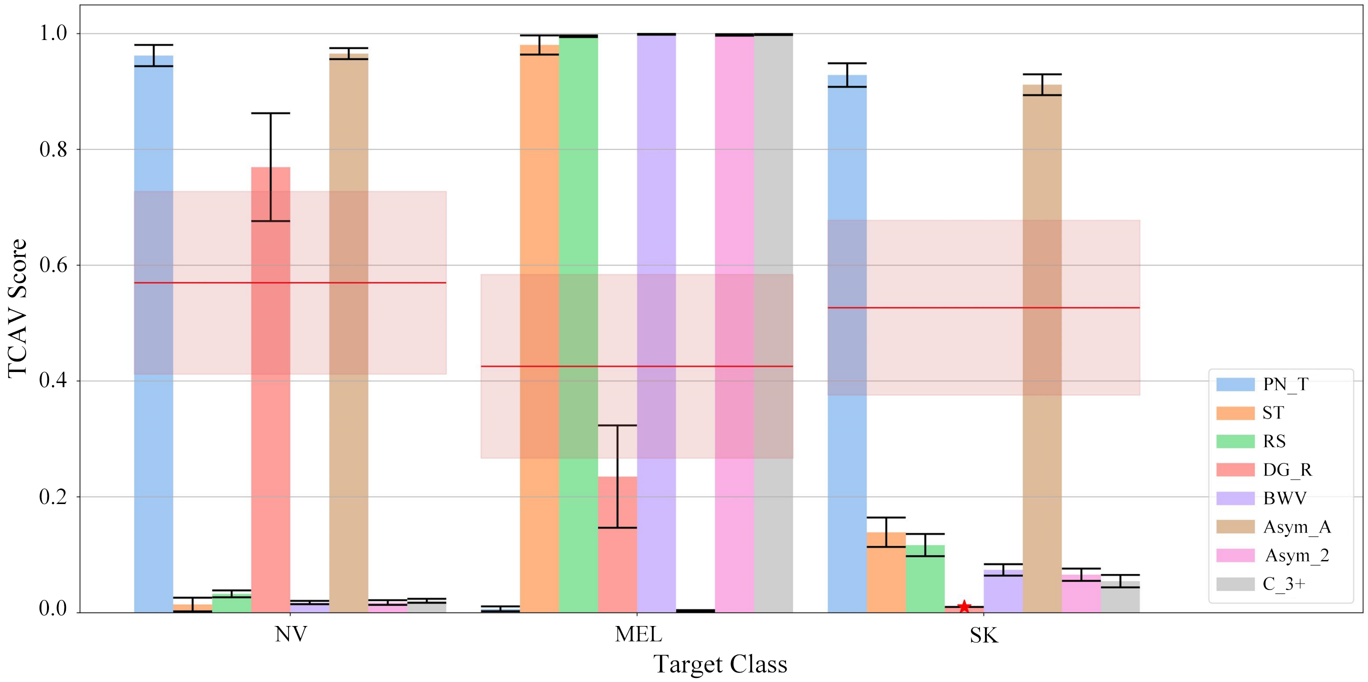


Fig. 5: TCAV scores of each concept for PH^2 with respect to each target class on the last layer of RECOD model.

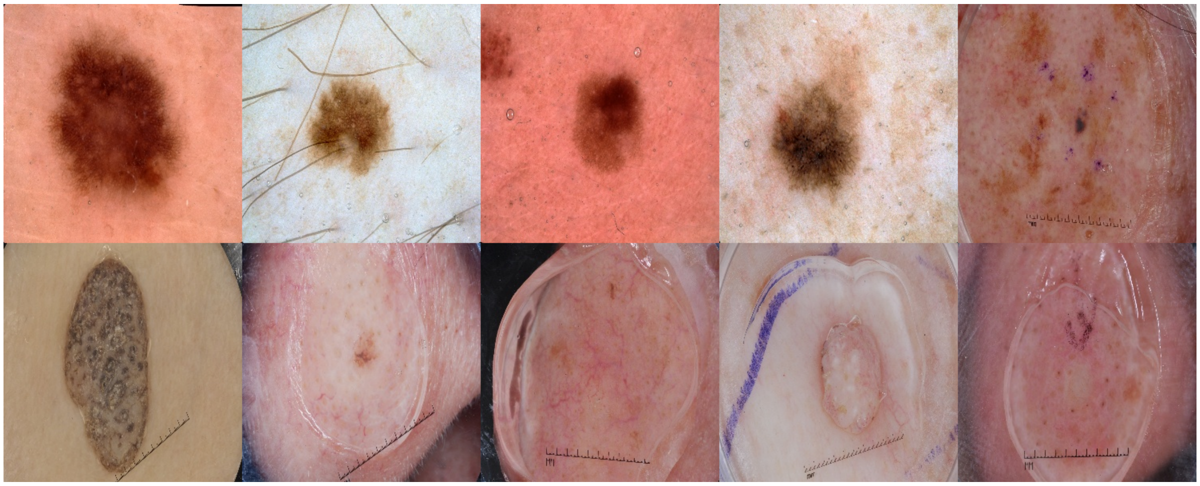


Fig. 6: First row shows the test images that are the most similar to Typical Pigment Network (*PN_T*) concept according to euclidean distance in CAV direction. The second row shows the least similar images to NP_T.

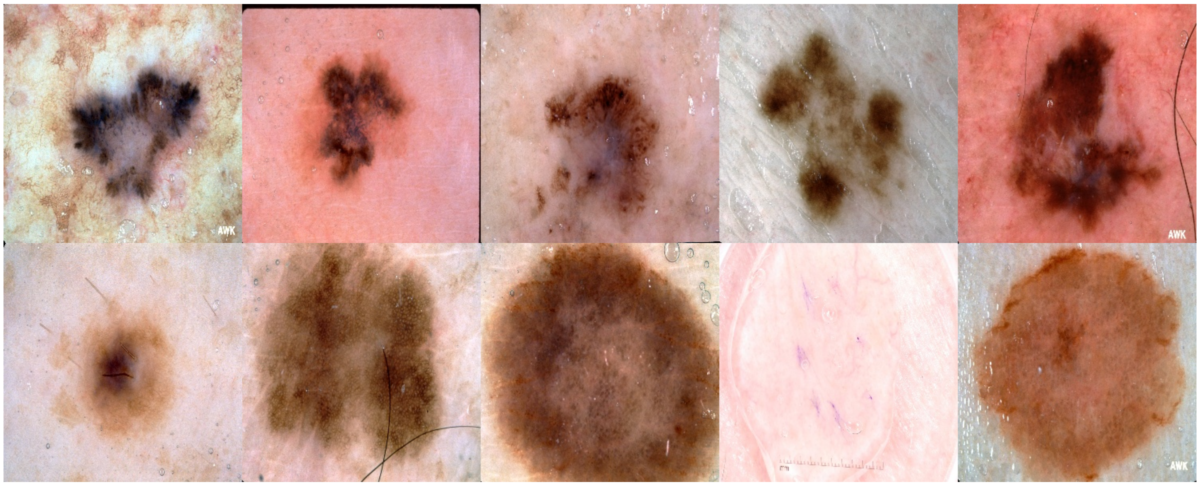


Fig. 7: First row shows the test images that are the most similar to Irregular Streaks (*ST_IR*) concept according to euclidean distance in CAV direction. The second row shows the least similar images to ST_IR.

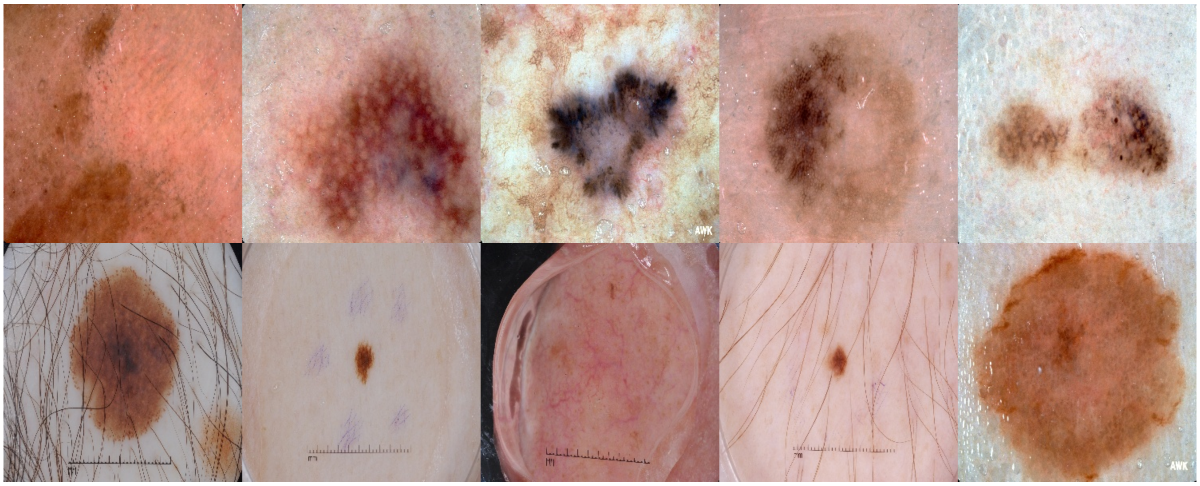


Fig. 8: First row shows the test images that are the most similar to Regression Structure (*RS*) concept according to euclidean distance in CAV direction. The second row shows the least similar images to RS.

systems that are able to justify their prediction by listing the concepts which influenced positively or negatively towards a certain output. It has also been shown that Testing with CAVs (TCAV) is applicable using complete IID images instead of general concept patches. However, this work can further be improved by using more granular labelling of diseases indicative concepts to get deeper insight into model's classification processes as well as further validation of its decisions.

Due to the complexity of the problem, possibly subjective annotations of training set by various expert, and small number of concept training samples, not all human-defined disease-related concepts were thoroughly analysed. Standardizing the annotation according to one *school of thought* in dermatology community, for example following [34], can decrease inter-observer disagreement but it would require an enormous amount of time and effort by dermatology expert. Nevertheless, the obtained results are remarkably aligned with common diagnostic criteria. In order to allow for a more comprehensive interpretation of the TCAV scores for this specific task, it would be desirable to curate a high-quality dataset with reliable fine-grained labels of concepts that are known to be highly indicative of specific diagnoses.

Supervised concept learning is highly dependent on high quality and precisely annotated human concept examples. Therefore, more focus should be placed on generating clean datasets of high-quality concept annotations that can be used for explaining models in medical imaging applications to speed up their deployments in clinical routines. As supervised concept classification from network activations has already been proven to be effective, the following logical and highly desirable extension of unsupervised concept discovery should be considered. Not only would this be another improvement towards simplifying the interpretability of networks by eliminating the necessity for laborious expert annotations but it could also allow insights in a network's own concepts, potentially revealing new knowledge for domain experts or unexpected biases in the network.

References

[1] United Nations General Assembly, "Transforming our World: The 2030 Agenda for Sustainable Development," available at https://sustainabledevelopment.un.org/post2015/, October 2015, a/RES/70/1.

[2] H. Zhao and M. Shingles, "AI for Good: Global Summit," June 2017.

[3] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, p. 115, 2017.

[4] M. N. Bajwa, M. I. Malik, S. A. Siddiqui, A. Dengel, F. Shafait, W. Neumeier, and S. Ahmed, "Two-stage framework for optic disc localization and glaucoma classification in retinal fundus images using deep learning," *BMC medical informatics and decision making*, vol. 19, no. 1, p. 136, 2019.

[5] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[6] Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," available at http://data.europa.eu/eli/reg/2016/679/2016-05-04, April 2016.

[7] A. B. Fortina, E. Peserico, A. Silletti, and E. Zattra, "Where's the naevus? inter-operator variability in the localization of melanocytic lesion border," *Skin Research and Technology*, vol. 18, no. 3, pp. 311–315, 2012.

[8] R. Corona, A. Mele, M. Amini, G. De Rosa, G. Coppola, P. Piccardi, M. Fucci, P. Pasquini, and T. Faraggiana, "Interobserver variability on the histopathologic diagnosis of cutaneous melanoma and other pigmented skin lesions." *Journal of clinical Oncology*, vol. 14, no. 4, pp. 1218–1223, 1996.

[9] A. Ghorbani, D. Ouyang, A. Abid, B. He, J. H. Chen, R. A. Harrington, D. H. Liang, E. A. Ashley, and J. Y. Zou, "Deep learning interpretation of echocardiograms," *bioRxiv*, p. 681676, 2019.

[10] S. Jolly, B. K. Iwana, R. Kuroki, and S. Uchida, "How do convolutional neural networks learn design?" in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 1085–1090.

[11] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that?" *arXiv preprint arXiv:1611.07450*, 2016.

[12] American Cancer Society, "Cancer facts & figures 2017," 2017, https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2017/cancer-facts-and-figures-2017.pdf.

[13] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA: a cancer journal for clinicians*, vol. 69, no. 1, pp. 7–34, 2019.

[14] E. S. Ruiz, F. C. Morgan, C. M. Zigler, R. J. Besaw, and C. D. Schmults, "Analysis of national skin cancer expenditures in the united states medicare population, 2013," *Journal of the American Academy of Dermatology*, vol. 80, no. 1, pp. 275–278, 2019.

[15] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," *arXiv preprint arXiv:1711.11279*, 2017.

[16] S. Palacio, J. Folz, J. Hees, F. Raue, D. Borth, and A. Dengel, "What do deep networks like to see?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3108–3117.

[17] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[18] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[19] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3319–3328.

[20] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[21] K. Young, G. Booth, B. Simpson, R. Dutton, and S. Shrapnel, "Deep neural network or dermatologist?" *arXiv preprint arXiv:1908.06612*, 2019.

[22] J. Wu, X. Li, E. Z. Chen, H. Jiang, X. Dong, and R. Rong, "What evidence does deep learning model use to classify skin lesions?" *arXiv preprint arXiv:1811.01051*, 2018.

[23] M. Munir, S. A. Siddiqui, F. Küsters, D. Mercier, A. Dengel, and S. Ahmed, "Tsxplain: Demystification of dnn decisions for time-series using natural language and statistical features," *arXiv preprint arXiv:1905.06175*, 2019.

[24] P. Guo, C. Anderson, K. Pearson, and R. Farrell, "Neural network interpretation via fine grained textual summarization," *arXiv preprint arXiv:1805.08969*, 2018.

[25] L. Anne Hendricks, R. Hu, T. Darrell, and Z. Akata, "Grounding visual explanations," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 264–279.

[26] B. Hancock, M. Bringmann, P. Varma, P. Liang, S. Wang, and C. Ré, "Training classifiers with natural language explanations," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2018. NIH Public Access, 2018, p. 1884.

[27] S. Srivastava, I. Labutov, and T. Mitchell, "Joint concept learning and semantic parsing from natural language explanations," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 1527–1536.

[28] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, "Mdnet: A semantically and visually interpretable medical image diagnosis network,"

in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6428–6436.

[29] B. Jing, P. Xie, and E. Xing, “On the automatic generation of medical imaging reports,” *arXiv preprint arXiv:1711.08195*, 2017.

[30] B. Zhou, Y. Sun, D. Bau, and A. Torralba, “Interpretable basis decomposition for visual explanation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 119–134.

[31] M. Graziani, V. Andrearczyk, and H. Müller, “Regression concept vectors for bidirectional explanations in histopathology,” in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Springer, 2018, pp. 124–132.

[32] M. Graziani, J. M. Brown, V. Andrearczyk, V. Yildiz, J. P. Campbell, D. Erdogmus, S. Ioannidis, M. F. Chiang, J. Kalpathy-Cramer, and H. Müller, “Improved interpretability for computer-aided severity assessment of retinopathy of prematurity,” in *Medical Imaging 2019: Computer-Aided Diagnosis*, vol. 10950. International Society for Optics and Photonics, 2019, p. 109501R.

[33] A. Ghorbani, J. Wexler, and B. Kim, “Automating interpretability: Discovering and testing visual concepts learned by neural networks,” *arXiv preprint arXiv:1902.03129*, 2019.

[34] H. Kittler, A. A. Marghoob, G. Argenziano, C. Carrera, C. Curiel-Lewandrowski, R. Hofmann-Wellenhof, J. Malvehy, S. Menzies, S. Puig, H. Rabinovitz *et al.*, “Standardization of terminology in dermoscopy/dermatoscopy: Results of the third consensus conference of the international society of dermoscopy,” *Journal of the American Academy of Dermatology*, vol. 74, no. 6, pp. 1093–1106, 2016.

[35] G. Argenziano, H. P. Soyer, S. Chimenti, R. Talamini, R. Corona, F. Sera, M. Binder, L. Cerroni, G. De Rosa, G. Ferrara *et al.*, “Dermoscopy of pigmented skin lesions: results of a consensus meeting via the internet,” *Journal of the American Academy of Dermatology*, vol. 48, no. 5, pp. 679–693, 2003.

[36] S. Menzies, C. Ingvar, and W. McCarthy, “A sensitivity and specificity analysis of the surface microscopy features of invasive melanoma.” *Melanoma research*, vol. 6, no. 1, pp. 55–62, 1996.

[37] R. J. Friedman, D. S. Rigel, and A. W. Kopf, “Early detection of malignant melanoma: the role of physician examination and self-examination of the skin,” *CA: a cancer journal for clinicians*, vol. 35, no. 3, pp. 130–151, 1985.

[38] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, “Ph 2-a dermoscopic image database for research and benchmarking,” in *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2013, pp. 5437–5440.

[39] A. Menegola, J. Tavares, M. Fornaciali, L. T. Li, S. Avila, and E. Valle, “Recod titans at isic challenge 2017,” *arXiv preprint arXiv:1703.04819*, 2017.

[40] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[41] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, “Seven-point checklist and skin lesion classification using multitask multimodal neural nets,” *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 538–546, mar 2019.

[42] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *et al.*, “Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic),” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172.

[43] R. P. Braun, H. S. Rabinovitz, J. Krischer, J. Kreusch, M. Oliviero, L. Naldi, A. W. Kopf, and J. H. Saurat, “Dermoscopy of pigmented seborrheic keratosis: a morphological study,” *Archives of dermatology*, vol. 138, no. 12, pp. 1556–1560, 2002.