ExAID: A Multimodal Explanation Framework for Computer-Aided Diagnosis of Skin Lesions

Adriano Lucieri^{*a,b,**}, Muhammad Naseer Bajwa^{*a,b*}, Stephan Alexander Braun^{*c,d*}, Muhammad Imran Malik^{*e*, f</sub>, Andreas Dengel^{*a*, *b*} and Sheraz Ahmed^{*a*}}

^aGerman Research Center for Artificial Intelligence (DFKI) GmbH, Trippstadter Strasse 122, 67663 Kaiserslautern, Germany

^bTechnical University Kaiserslautern, Erwin-Schrödinger-Strasse 52, 67663 Kaiserslautern, Germany

^cUniversity Hospital Münster, 48149 Münster, Germany

^dUniversity Hospital of Düsseldorf, 40225 Düsseldorf, Germany

^eSchool of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan ^fDeep Learning Laboratory, National Center of Artificial Intelligence, Islamabad, Pakistan

ARTICLE INFO

Keywords: Artificial Intelligence in Dermatology Computer-Aided Diagnosis Explainable Artificial Intelligence Interpretability Medical Image Processing Textual Explanations

ABSTRACT

Background and Objectives: One principal impediment in successful deployment of Artificial Intelligence (AI)-based Computer-Aided Diagnosis (CAD) systems in everyday clinical workflow is their lack of transparent decision making. Although commonly used eXplainable AI (XAI) methods provide some insight into these largely opaque algorithms, yet such explanations are usually convoluted and not readily comprehensible except by highly trained AI experts. The explanation of decisions regarding the malignancy of skin lesions from dermoscopic images demands particular clarity, as the underlying medical problem definition is itself ambiguous. This work presents and evaluates ExAID (Explainable AI for Dermatology), a novel XAI framework for biomedical image analysis, providing multi-modal concept-based explanations consisting of easy-to-understand textual explanations supplemented by visual maps justifying the predictions.

Methods: Our framework relies on Concept Activation Vectors (CAVs) to map humanunderstandable concepts to those learnt by an arbitrary Deep Learning based algorithm in its latent space, and Concept Localisation Maps (CLMs) to highlight concepts in the input space. This identification of relevant concepts is then used to construct fine-grained textual explanations supplemented by concept-wise location information to provide comprehensive and coherent multi-modal explanations. All decision-related information is comprehensively presented in a diagnostic interface for use in clinical routines. Moreover, the framework includes an educational mode providing dataset-level explanation statistics and tools for data and model exploration to aid medical research and education processes

Results: Through rigorous quantitative and qualitative evaluation of our framework on a range of dermoscopic image datasets such as SkinL2, Derm7pt, PH² and ISIC, we show the utility of multimodal explanations for CAD-assisted scenarios even in case of wrong disease predictions.

Conclusions: We present a new multi-modal explanation framework for biomedical image analysis on the example use-case of Melanoma classification from dermoscopic images and evaluate its utility on a row of datasets. Since comprehensible explanation is one of the cornerstones of any CAD system, we believe that ExAID will provide dermatologists an effective screening tool that they both understand and trust. Moreover, ExAID will be the basis for similar applications in other biomedical imaging fields

1. Introduction

arXiv:2201.01249v1 [cs.AI] 4 Jan 2022

In 2016, Ribeiro et al. [36] reported an image classifier that was able to inadvertently classify correctly but for wrong reasons. They found out that their wolf versus dog classifier learnt an undesirable correlation between wolf and background snow and, therefore, would classify a given image as wolf if there was snow in the background. If it were not due to authors' vigilance in finding explanations to the model's predictions, it would have been difficult to properly evaluate trustworthiness of this image classifier. Although this was an inconsequential example of spurious correlations learnt from a large amount of data, wrong decisions in safety

Email addresses: adriano.lucieri@dfki.de (A. Lucieri); naseer.bajwa@dfki.de (M.N. Bajwa); stephanalexander.braun@ukmuenster.de (S.A. Braun);

malik.imran@seecs.edu.pk (M.I. Malik); andreas.dengel@dfki.de (A. Dengel); sheraz.ahmed@dfki.de (S. Ahmed)

critical domains resulting from such misunderstandings can potentially have grave impact on human lives. The application of Artificial Intelligence (AI) methods on medical tasks has become ubiquitous in the last decade [1, 24, 39]. Therefore, hesitation of medical practitioners in trusting diagnostic predictions of any automated Computer-Aided Diagnosis (CAD) system is understandable since such systems normally provide little to no cognisance regarding their decision making process [28]. In addition to evaluating the reasons behind a model's predictions, explanation methods can also help in revealing new diagnostic criteria [21] previously unknown to medical practitioners. The requirement for a CAD to be explainable arose with early applications of AI in healthcare and became more relevant with recent ethical and legal standards [3, 11]. The consequent increase in research activity in the domain of eXplainable AI (XAI) also reflects the growing interest of the community to provide ex-

^{*}Corresponding author

planations for CAD systems [3].

XAI methods for the image-domain come in a variety of forms and provide explanations using a range of modalities such as feature-relevance visualisations [36, 15, 43, 6], textual explanations [20, 46], or quantitative relevance measures for abstract concepts [25, 7]. They differ not only in the way they are presented to their users but also in their derivation, resulting in varying levels of insight provided regarding the decision making of the AI. Furthermore, these methods can be either ante-hoc (e.g. ProtoPNet [6], MDNet [46]), with a decision making process that is explainable by design, or post-hoc, providing explanations for an AI model after construction and training of the model using model-specific (e.g. Score-CAM [43], TCAV [25]) or model-agnostic (e.g. LIME [36], EP [15]) techniques [34]. Most methods provide explanations on a local scale (individual data samples) while some aim at approximating explanations on a global scale (holistic model behaviour). However, model explanations given by single XAI methods are usually not sufficient to provide plausible and easy-to-understand decision justification to end users.

Melanoma is the most dangerous skin cancer, leading to the majority of skin-related deaths in the US while accounting for only 1% of skin cancers diagnosed [37]. Regular preventive examinations are conducted by physicians through naked-eye observation or dermoscopic imaging. In dermoscopic pattern recognition, experts look for dermoscopic criteria and apply manual algorithms like the ABCDrule [35] or 7-point checklist [2] to judge the malignancy of a lesion. Currently, AI-based dermatology focuses mostly on the analysis of dermoscopic images [31, 41, 32]. However, first approaches towards analysing raw, clinical images have been proposed as well [5]. The majority of explanation approaches for dermoscopic skin lesion analysis rely on the application of visual XAI through saliency maps [44, 45] or attention mechanisms [17, 4]. Another common approach is the detection and localisation of dermoscopic criteria as used by doctors in manual classification. Coppola et al. [10], for instance, train a multi-task CNN predicting dermoscopic features with information sharing between different subnetworks to increase interpretability. In [27], Lucieri et al. apply the concept-based TCAV method to predict dermoscopic criteria from a pre-trained network to explain its predictions. Dermoscocpic criteria localisation has been been approached by combining TCAV with perturbation-based saliency methods in [29] and through explicit segmentation of criteria in [38]. For a complete survey on XAI in Dermatology the reader is referred to [30].

Several frameworks for AI-based medical imaging have been proposed in recent years [16, 38, 19, 22]. While some lack proper and comprehensible explainability, others do not provide an easy-to-use interface for human-machine interaction, impeding the utilisation in diagnostic routines or research. Moreover, first commercial platforms for biomedical AI exist [40, 12, 13, 18], claiming to provide explanations for their algorithms.

In this paper we present a novel XAI framework, namely

Explainable Artificial Intelligence for Dermatology $(ExAID)^{1}$ which is able to provide easy-to-understand textual, visual and conceptual explanations for automated analysis of dermoscopic images of malignant and benign skin lesions, while being adaptable to any other biomedical imaging use case. ExAID is built upon two of our previous works: [27], which verifies that deep learning models are able to learn and utilise similar disease-related concepts as described by dermatologists and employed by them during manual analysis of skin images; and [29], which localises these concepts, learnt and embedded in the latent space of the model, on the original image. ExAID extends the previously proposed explanation modalities by introducing concept-based textual explanations while integrating all modalities in a unified, intuitive framework to further enhance intelligibility of AI's decision making in a diagnostic setting. In addition to clinical diagnosis functionalities, it provides in-depth analysis tools for medical research and education. Therefore, the framework offers two distinct interfaces for clinical diagnosis and research purposes, laying the foundation for understandable and transparent integration of AI in medical workflows. In contrast to existing XAI frameworks, ExAID emphasises intuitive intelligibility for end users by conveying multi-modal decision justification centred around standard concepts commonly used in the dermatology domain.

The rest of the paper is organised as follows.

Section 2 covers details of the datasets used for training and evaluation as well as an introduction to the framework and its components. Experimental setups for the generation of explanations and their evaluation are described in Section 3. Finally, Section 4 discusses the presented results and elaborates on current limitations of this particular use case and framework state, before the work is concluded in Section 5.

2. Methods

2.1. Datasets

ExAID contains two types of classifiers: Disease-level classifiers for lesion diagnosis and concept-level classifiers for detection of dermatological concepts in a given image. To train these two classifiers, it requires datasets with two types of labels, namely disease labels (*Melanoma* and *Nevus*) and concept annotations (presence or absence of dermoscopic concepts).

2.1.1. Datasets for Disease-level Classification

The training set for disease-level classification consists of *Melanoma* and *Nevi* images taken from ISIC 2019 as well as PH² [33] and derm7pt [23] datasets. ISIC 2019 dataset is a public collection of 25,331 images of different provenance divided into eight different classes. This dataset is a coalition of three datasets, HAM10000 [42], BCN20000 [9], and MSK [8]. Since the common denominator of ISIC2019, PH², and derm7pt datasets are *Melanoma* and *Nevi* classes, we assembled a subset of the three datasets consisting of im-

¹A demo will be soon available under https://exaid.kl.dfki.de/.

Table 1

Distribution of data in training, validation and test splits for disease-level classification.

Solit	Dataset	Lesions				
Spin	Dataset	Melanoma	Nevi	Total		
	ISIC2019	1250	2894	4144		
Train	Derm7pt	158	368	526		
	PH2	26	102	128		
	ISIC2019	313	723	1036		
Validate	Derm7pt	40	92	132		
	PH2	6	26	32		
	ISIC2019	391	904	1295		
Test	Derm7pt	50	115	165		
	PH2	8	32	40		
	ISIC2019	1954	4521	6475		
Total	Derm7pt	248	575	823		
	PH2	40	160	200		

Table 2

Distribution of data in training, validation and test splits for concept-level classification with D7PH2 dataset.

Salit	Dotocot	Lesions			
Shirt	Dalasel	Melanoma	Nevi	Total	
Train	Derm7pt	158	368	526	
ITalli	PH2	26	102	128	
Validate	Derm7pt	40	92	132	
	PH2	6	26	32	
Tost	Derm7pt	50	115	165	
Test	PH2	8	32	40	
Total	Derm7pt	248	575	823	
	PH2	40	160	200	

ages from these two classes only, and manually cleansed the dataset for duplicates and samples with low quality (e.g. systematic artefacts), resulting in a total of 6,475 images. As PH² and derm7pt will be used for training the concept-level classifiers, a custom dataset split is assembled from a combination of all three stratified datasets to avoid covariate shifts between disease-level and concept-level training stages. The distribution of images in training, validation and test sets for disease-level classification is given in Table 1. The generalisability of the model is moreover evaluated on a range of other datasets including 2016 and 2017 ISIC challenge datasets and SKINL2 [14] dataset as shown in Table 3.

2.1.2. Datasets for Concept-level Classification

Training of concept classifiers requires annotations regarding presence or absence of specific dermoscopic concepts. These annotations are not usually available with dermoscopic image datasets, which limits our selection of training and evaluation datasets primarily to PH2 and derm7pt. The PH² dataset is a small dataset of only 200 dermoscopic images containing 80 common nevi, 80 atypical nevi, and 40 melanoma. For each image, the dataset provides colour and lesion segmentation masks and extensive, well-curated annotations with respect to presence or absence of various concepts. The derm7pt dataset contains 1,011 clinical and dermoscopic images divided into four diagnosis classes and one miscellaneous class. Two of these diagnosis classes, *Melanoma* and *Nevi* are further divided into 13 sub-classes. From this dataset, 823 images belonging to only *Melanoma* and *Nevi* samples have been considered. The combination of derm7pt and PH² used for concept classification is subsequently referred to as D7PH2. Table 2 shows the distribution of images used in the concept-level classification task. ISIC 2016 and 2017 challenge datasets are moreover used for the evaluation of concept classifier generalisability. However, both datasets only include annotations of two dermoscopic concepts each, namely *Pigment Networks* and *Streaks* as well as *Dots & Globules* and *Streaks*, respectively.

2.2. ExAID Framework

At its core, ExAID is a generic toolbox for human-centred post-hoc explanations able to explain arbitrary DL-based models even beyond applications in dermatology. In addition to the DL model to be explained, its computational foundation consists of three basic components, namely Concept Identification, Concept Localisation and Decision Explanation modules as depicted in Fig. 1.

2.2.1. Concept Identifier

The Concept Identifier maps disease-related dermatological concepts to their corresponding representation learnt by the DL-based model in its latent space using Concept Activation Vectors [25] (CAVs). For each pre-defined concept a linear binary classifier is trained on the detection of said concept from the model's activation space, resulting in a CAV which represents the main concept direction in this latent space. CAV training can be executed on arbitrary model layers, automatically selecting each concept's best performing activations for inference.

Once trained, the concept classifiers allow to predict presence or absence of single concepts on unseen images, based on the model's latent activation patterns. CAVs additionally allow for computation of the global TCAV metric, estimating a concept's overall contribution to the prediction of a certain target class. Further details on the Concept Identifier module and the CAV training procedure can be found in [27].

2.2.2. Concept Localiser

Concept Localisation Maps (CLMs) [29] extend CAVs by localising regions pertinent to a learned concept in the latent space of a trained image classifier. They provide qualitative and quantitative assurance of the model's ability to learn the right interpretation of a concept by indicating the exact spatial location that contributed to a concept prediction and moreover enable the visualisation of other, potentially abstract concepts.

Given an input image $x \in X$, a linear concept classifier g_C generates a concept score for a concept *C* based on the trained model's latent vector $f_l(x; \theta)$ at layer *l* with optimal weights θ . The Concept Localiser implements the perturbation-based concept localisation technique from [29] to generate spatial importance values based on variation of



Figure 1: ExAID Framework architecture. The schematic drawing shows the in-, output and flow of information through ExAID as well as the relationship between its components.

the concept scores $g_C(f_l(x; \theta))$. The resulting map m_{Cl} corresponds to the input region contributing most to concept *C*. Instead of occlusion by means of black patches, in this work a radial Gaussian mask is applied to a blurred image patch to mitigate distribution shift in perturbed images stemming from sharp edges and colour gradients in the perturbed images.

2.2.3. Decision Explainer

The Decision Explainer receives all concept prediction scores for a given image from the Concept Identifier. A rulebase is derived from a calibration dataset and applied to the translation of single concept scores into a textual decision explanation grounded in human-understandable conceptual evidence. An explanation sentence conveys graded information about the conceptual evidence detected by a given model, as well as its influence to the given prediction. An example for a textual explanation along with the corresponding input image is given in Fig. 2.

The explanations derived from concept detection are composed into coherent and easy-to-understand explanation texts. An explanation sentence is constructed based on concept scores and directional derivatives computed during concept detection under discrimination between absence, moderate evidence and strong evidence of concepts to reflect the fuzzy nature of concepts' appearance. Manifestation of a concept is decided by means of thresholds derived from the concept training data. This is achieved by first scaling the unbound concept prediction using a two-sided normalisation scheme to obtain a centred probability of concept presence. Thresholds are then derived by maximising False Positive Rate (FPR) and True Positive Rate (TPR) among all positive predictions on the training dataset for moderate and strong evidence thresholds, respectively. The directional derivatives of the predicted class along the individual CAV is used to indicate positive or negative influence of concept to the prediction. Conceptual evidence is listed after the key word "despite" in case of negative class influence to signalise contraindication (see Figure 5).

2.3. Operation Modes

ExAID offers two complementary operation modes that are meant for different use cases. A diagnostic mode provides functionality meant to support dermatologists during clinical examination of patient's skin lesions. For research and education purposes, ExAID offers an educational mode including a collection of tools for holistic analysis of the deep model's behaviour as well as the collected case data.

2.3.1. Diagnostic Mode

The majority of a dermatologist's clinical routine consists of visual examination of patients' skin lesions to reach a decision regarding further investigation of a potentially malignant lesion. Provided enough evidence for malignancy,



Figure 2: Diagnostic Mode of ExAID can be used as Decision Support System in routine clinical workflow.

the suspicious tissue may be excised under local anaesthesia. Physicians with considerable experience in dermoscopy develop an intuition which enables them to promptly reach to a conclusion while novices initially need to pay greater attention to the assessment of a particular skin lesion. This is among other things owed to the disarray of dermoscopic terms and concepts and their usage in different schools of thought. Having developed a routine and diagnostic intuition not only bears the risk of subjective bias in a decision, but it might also lead to negligence in the identification of important diagnostic details, which is furthermore aggravated by emotional stress and time constraints.

ExAID's diagnostic mode aims at mediating subjectivity by offering a supplement to the experienced physician's first impressions, serving as a second opinion which stimulates physician's thought and breaks the routine. Through this additional information it is made sure that cues vital for successful identification of malignant conditions are not overseen during manual examination. The user interface of the diagnostic mode is presented in Fig 2. While allowing physicians to examine the dermoscopic image manually, an initial diagnosis suggestion is provided, supported by conceptbased textual, quantitative and visual explanations. Through its neutral design, the interface assures that users are not biased towards the proposed diagnosis but are free to reconstruct the AI's decision process by considering and validating biomarker scores along with their optional localisations provided in the form of CLMs.

2.3.2. Educational Mode

Explanation of classifiers' decisions has further utility beyond mere information and guidance of the algorithm's users. It is of central importance for the validation of individual automated decisions, verification of plausibility of a model's global generalisation behaviour and can additionally aid the decryption of unintelligible, decision relevant concepts learned by the AI. With its educational mode, as presented in Fig. 3, ExAID offers an extensive toolbox for the investigation of model behaviour and data distribution. Dataset-level model behaviour analysis is enabled through a combination of class-wise performance evaluation metrics and concept-wise global explanation metrics in combination with tools for facilitated overview of individual decision outcomes and explanations. Some of the most salient interactive features of ExAID framework are introduced below.

Filtering The filtering option allows to filter arbitrary subsets of samples by metadata such as age, concept presence, concept prediction or correct prediction. An adaptive data distribution plot helps to quickly identify important statistical characteristics related to biomarker presence as well as certain failure modes of the model.

Highlighting A highlighting feature allows to spotlight certain useful properties of samples to further facilitate review of model behaviour and data. This feature allows the highlighting of not only binary attributes such as the correct target class prediction, but also more complex relationships such as the presence of classes or concepts in the annotations as



Figure 3: Educational Mode of ExAID can help in training of resident dermatologists by allowing them to explore many of its interactive features.

well as the class and concept prediction by the model. Complex highlighting is always supported by visual cues indicating the accordance of attribute prediction with expert annotations.

Localisation In addition to individual localisation of concepts in data samples, ExAID allows to visualise concept localisation simultaneously for all samples of a dataset. This allows for quick examination of a model's concept localisation behaviour, aiding the validation of system behaviour and identification of potential systematic errors in dataset or model by revealing patterns in the localisation process.

Latent Inspection Examination of the model's latent space structure gives further insight into the disentanglement of data representations and potential biases captured by the model parameters. A latent view functionality based on Tensorboard's projector² allows to intuitively examine the latent distribution of data samples by means of dimensionality reduction techniques.

3. Results

3.1. Classifier Training & Evaluation

To demonstrate the utility of the proposed framework, a deep network for binary classification of *Melanoma* and *Nevi* from dermoscopic skin lesion images is trained. Among various architecture, learning rate and optimiser combinations³,

Table 3 Performance evaluation of lesion classifier on various datasets

٢	rerformance	evaluation	ot	lesion	classifier	on	various	datasets	;
			_						

Datasets	Ν	Accuracy (%)	Precision (%)	Recall (%)	AUC
Derm7pt (Test)	165	83.6	81.7	78.0	0.85
PH2 (Test)	40	100.0	100.0	100.0	1.00
ISIC2019 (Test)	1295	88.9	88.2	84.9	0.91
ISIC2017 (Test)	510	78.4	68.5	62.3	0.70
ISIC2016 (Test)	379	89.7	83.7	84.0	0.92
SkinL2	55	90.9	89.9	90.7	0.99

best results have been achieved using SEResNeXt architecture with RMSprop optimiser and a learning rate of 1e-4 trained for 100 epochs. Training images were augmented by random horizontal and vertical flip as well as random cropping to 85% of the image size, resulting in input images of size 224×224 .

Evaluation on a variety of datasets is presented in Table 3. It can be observed that the lesion classifier achieved AUCs above 0.85 for five out of six datasets with two datasets scoring almost perfectly. ISIC2017 achieved a slightly lower AUC with 0.70, which is also reflected in lower Precision and Recall.

3.2. Explanation Training & Evaluation

For the explanation of the final DL-based classifier's decisions, the procedure outlined in [27] is followed. To this end, concept annotated samples from D7PH2 have been utilised to assure generalisation while learning CAVs. In each run, the data is internally split into folds for concept training and validation under stratification of both concept and disease la-

²https://projector.tensorflow.org/

³Experimentation included VGG16, ResNet, DenseNet, NASNet, SEResNeXt architectures with Adam, SGD and RMSprop optimisers using learning rates ranging from 1e-3 to 1e-4



(a) Streaks.

(b) Pigment Network.

Figure 4: Positive and negative examples of visual explanations provided by ExAID along with the corresponding samples and ground truth concept masks.

Table 4

Performance evaluation of concept classifiers on various datasets. Results are given as Macro Average F1-Scores to account for class imbalance.

Datasets	Streaks	Pigment Netw.	Dots & Glob.	Regr. Struct.	Blue-Whit. Veils
D7PH2 (Test)	70.91	78.74	63.14	59.41	71.66
ISIC2017	51.75	50.37	-	-	-
ISIC2016	56.53	-	53.03	-	-

bels. For each concept, linear concept classifiers are trained for 200 runs using stochastic gradient descent with early stopping.

3.2.1. Concept Detection

The final CAV for a concept is chosen based on the average concept direction over all runs. Due to concept annotation requirements, concept detection performance is evaluated only on ISIC2016 and ISIC2017 datasets as well as D7PH2 test set. Due to the lack of annotation, the two ISIC datasets allowed evaluation of only two concepts each. Table 4 presents Macro Average F1-Scores for concept detection.

For D7PH2 test set, all concept detectors were able to discriminate concepts better than random guessing, with *Pigment Network* achieving the best F1-Score of 78.74%. Same holds for the evaluation on ISIC2016 for concepts *Streaks* and *Dots & Globules*. Similar to the results for lesion classification on ISIC2017, concept detectors failed to classify *Streaks* and *Pigment Networks*, yielding F1-Scores of 51.75% and 50.37%, respectively.

3.2.2. Visual Explanation

Fair quantitative evaluation of a network's CLMs for skin lesions poses a number of difficulties including the selection of a suitable binarization scheme, subjectivity of concept annotations as well as lack of representative metrics for fuzzy localisation tasks. Proper binarization is specially difficult as it depends on the size of a particular Region of Interest (ROI), its significance to the prediction score as well as further noise stemming from the saliency method used. Moreover, evaluation is limited by the availability of annotated concept segmentation maps. ISIC2016 and ISIC2017 challenge datasets each provide concept segmentation maps for two concepts which are used to provide a qualitative assessment of the trained model's concept localisation ability. CLMs were binarized using variable percentiles, manually chosen based on the size of the respective ROI in a specific image. Figure 4 shows examples of the model's concept localisation ability for classes *Streaks* and *Pigment Network* using an adaptation of the method proposed in [29]. Interpretation of the results is provided in Section 4.3.

3.2.3. Textual Explanation

Quantitative evaluation of textual explanation results is covered by the performance evaluation for concept detection presented in table 4. Figure 5 shows qualitative examples of images along with correct and incorrect textual explanations provided by ExAID. These results are further discussed in Section 4.4.

4. Discussion

The results presented in Section 3 show that ExAID is indeed able to produce theoretically correct explanations for a classifier's decisions. In the following, previous insights are discussed and qualitative results are analysed in detail. Finally, a detailed discussion about the limitation and future work is provided.

4.1. Lesion Classification

Lesion-level results clearly show the strong generalisation ability of the model, even on unseen datasets as SKINL2 consisting of 20 Melanomas and 35 Nevi of high quality. Poor performance on the ISIC2017 test dataset can be explained by the large fraction of artefacts present in the images, which have been intentionally left out of the training

ExAID: A Multimodal XAI Framework for CAD of Skin Lesions



The image has been identified as Melanoma due to moderate evidence of Irregular Dots & Globules, and strong presence of Blue-Whitish Veil, and absence of Pigment Networks and Regression Structures.

The image has been identified as Nevus due to

and absence of Atypical Pigment Networks,

Regression Structures and Blue-Whitish Veil.

The image has been identified as Nevus due to absence of Atypical Pigment Network

and strong evidence of Irregular Dots & Globules,

Melanoma



Nevus



Melanoma

(a) Correct concept prediction.

and Regression Structures,

despite strong evidence of Streaks,

and strong evidence of Blue-Whitish Veil.

moderate evidence of Streaks,



The image has been identified as Nevus due to moderate evidence of Typical Pigment Networks, and absence of Atypical Pigment Network and Blue-Whitish Veil, despite moderate evidence of Regression Structures, and moderate evidence of Irregular



The image has been identified as Nevus due to moderate evidence of Streaks, despite moderate evidence of Dots & Globules.



The image has been identified as Melanoma due to moderate evidence of Blue-Whitish Veil, despite moderate evidence of Regular Dots & Globules

Dots & Globules.

(b) Incorrect concept prediction.

Figure 5: Positive and negative examples of textual explanations provided by ExAID along with the corresponding skin lesion samples. The ground truth class of the sample is given below the image.

procedure (through manual cleansing) to restrict the use case to a realistic, controlled diagnostic environment based on an image acquisition procedure specifically built for AI processing.

4.2. Concept Detection

In contrast to the concept detection performance on the D7PH2 test set, concept generalisation to unseen datasets such as ISIC2017 and ISIC2016 is worse. This is most likely a consequence of diverging annotation standards between derm7pt and PH² datasets used for CAV training and other datasets. The distribution shift partially caused by artefacts present in the challenge test sets aggravates this divergence further. Moreover, results show the superiority of coarse-grained biomarkers such as *Streaks*, *Pigment Networks* and *Blue-Whitish Veils* over more fine-grained ones such as *Dots & Globules*.

4.3. Visual Explanation

Whereas in some cases, CLM localisation aligned very well with the concept annotation, most of the time CLMs highlighted slightly different regions. However, these highlights often depict areas that could plausibly count as concept regions, as can be seen in the second row of Figure 4. Qualitative evaluation confirmed the quantitative results and showed that the network performed better localising concepts *Streaks* and *Pigment Networks* as compared to the more fine-grained *Dots & Globules* concept. Scattered spots in CLMs outside the lesion regions highlights noise problems inherent in perturbation-based CLM computation and the dependence on a proper binarization scheme.

4.4. Textual Explanation

Examples in Figure 5a depict instances with correct concept predictions, which showcase the simplicity and intelligibility of the generated explanations. Explanation texts briefly reflect the most important criteria necessary for experts to understand the network's decision. Interestingly, it appeared that although correct concept predictions where given, the network sometimes misclassified the underlying disease as seen in the third row of Figure 5a. This could be due to the presentation of an ambiguous borderline case or a result of wrong ground truth annotation for either lesion class or concepts. However, the explanation explicitly exposes Streaks, Irregular Dots & Globules and Blue-Whitish Veil as contraindications for the prediction of Nevus. In a clinical setting, such contraindication would raise the suspicion of a user, possibly initiating a more thorough review of the case. This particularly emphasises the utility of such a system, as a correct explanation will allow physicians to scrutinise a given prediction, not solely relying on an automated, opaque categorical output value.

Figure 5b on the contrary shows failure cases where the network confused different visual cues for concepts. While *Irregular Dots & Globules* have been correctly detected in the top right image, the middle right image contains white blobs which might have been confused as *Dots & Globules* by the model. The bottom right case shows a *Blue Nevus* which has been confused by the network as a *Melanoma* showing signs of *Blue-Whitish Veil* although containing *Regular Dots & Globules*. It is notable that samples with incorrect concept predictions already expose a certain uncertainty by exhibiting moderate concept detections more frequently as compared to the samples from Figure 5a as well as their prevalence of contraindications. This clearly shows that irrespective of the model used for prediction, ExAID is able

to provide well-founded justifications which help to express model uncertainty, encouraging closer examination of rare and edge cases.

4.5. Limitations

This proof-of-principle study primarily focuses on the current state of the proposed framework with its comprehensible user interface, conveying textual, visual and conceptual explanations for trustworthy computer-aided decision support in medicine. The development of the framework is an ongoing Co-Design process which holistically includes a wide variety of stakeholders with different needs to assure not only practical but also ethical aspects early on [47]. Although, concept classification, localisation and textual explanation abilities of ExAID are remarkable given the fact that the DL model has not explicitly been trained on those tasks, some challenges must be first solved before an application in real clinical settings becomes feasible.

Current public datasets often suffer from the low sample quality attributable to a lack of process standardisation⁴, missing histological diagnosis confirmation and subjective annotation. Together with the low number of overall available images, in particular the ones with detailed concept annotation, this results in significant shift of data distributions between different datasets, constituting the major reason for sub-optimal generalisation of the proposed concept classifiers to other datasets.

ExAID's concept localisation ability yet suffers from limitations due to the perturbation-based nature of saliency map generation which results in noisy heatmaps and high sensitivity to hyperparameters, specially in case of varying sizes of biomarkers. Future work applying optimisation-based perturbation methods for concept localisation will mitigate those issues, resulting in more flexible and robust heatmaps. Textual explanations are generated based on concept predictions as well as directional derivatives as used in TCAV scores. Lacking a meaningful scaling of gradients, only the direction and not the magnitude of a concept's influence is currently used to improve the explanation text. Incorporation of more robust concept influence measures could add another level of details to the rule-base, making the explanations more differentiated and rendering the system even more useful in practice.

Quantitative evaluation of concept detection or localisation is still limited due to lack of similarly and sufficiently annotated data from other sources. To solve this issue, an agreed upon definition and consensual annotation of a large number of representative images is required, which will reflect in higher quality explanations. Moreover, evaluation of CLMs is aggravated by noise artefacts emerging during binarisation and lack of definite measures for fuzzy localisation tasks. A qualitative evaluation in real-world setting by medical experts is of extreme value for the evaluation of the explanations' utility to the diagnostic workflow and will be realised as soon as may be.

The influence of subjectivity not only reflects in the data annotations, but also in the general uncertainty surrounding the field of dermoscopy. Despite first attempts towards standardisation of dermoscopic terminologies and concepts [26], no general consensus has yet been broadly established among physicians. Thus, a variety of diagnostic schools prevail and interpretation of terms and concepts is still largely depending on the education, preference and experience of the individual physician. This work focuses on the 7-point checklist criteria [2] as well as further dermoscopic concepts from [33], due to the public availability of annotated data. The commitment to a specific set of concepts prior to the decision of a standard consensus might hamper the acceptance of the framework by physicians accustomed to different methods and the mixture of different schools and interpretations of concepts bears the risk of contrasting labelling. Productive deployment of such a system requires diligent assessment through medical practitioners in real-world environments, providing their valuable feedback to evaluate and improve such a system. Prior to performing clinical trials, the system should be fed with carefully selected data properly representing a set of meaningful and unambiguously defined dermoscopic concepts as agreed by a committee of dermatology experts.

5. Conclusion

Since the advent of modern DL-based systems and their industrious applications in medical domains, there have been remarkable strides in the explanation of these complex systems that, in some cases, already led to correction and verification of AI as well as disclosure of new potential diagnostic criteria. However, sensible and comprehensible explanations is still one of the greatest challenges related to medical image diagnosis, which should be addressed by concerted efforts from AI researchers, medical practitioners and regulatory authorities.

With ExAID, this article presents a framework which consolidates and builds upon our previous works on detection of human-defined concepts for skin lesion diagnosis in DL model's latent space and their localisation on the input image to provide intelligible textual explanation of model's predictions. We showed that, despite severe limitation in terms of data and annotation availability, the system already provides useful insights into DL classifier's decision making, even in case of wrong predictions. When properly addressing the current limitations, this framework will not only play a useful assistive role in reliable, efficient and objective screening of melanoma, which is one of the most serious skin cancers, but also help train new dermatologists efficiently and effectively.

The generality of the framework allows its adaptation to various other image-based domains like radiology or histology. In view of future deployment of the system in clinical practice, its under-lying general purpose DL architecture will be further specialised to the respective domain, includ-

⁴Different camera setups, operators and techniques like polarised and non-polarised dermoscopy resulting in varying image quality, lighting, alignment and artefacts.

ing improvements to the processing pipeline such as lesion segmentation and hair removal.

Acknowledgement

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 and BMBF projects ExplAINN (01IS19074) and DeFuseNN (01IW17002).

References

- Al-Antari, M.A., Han, S.M., Kim, T.S., 2020. Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital x-ray mammograms. Computer Methods and Programs in Biomedicine 196, 105584.
- [2] Argenziano, G., Fabbrocini, G., Carli, P., De Giorgi, V., Sammarco, E., Delfino, M., 1998. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the abcd rule of dermatoscopy and a new 7-point checklist based on pattern analysis. Archives of dermatology 134, 1563–1570.
- [3] Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58, 82–115.
- [4] Barata, C., Celebi, M.E., Marques, J.S., 2021. Explainable skin lesion diagnosis using taxonomies. Pattern Recognition 110, 107413.
- [5] Birkenfeld, J.S., Tucker-Schwartz, J.M., Soenksen, L.R., Avilés-Izquierdo, J.A., Marti-Fuster, B., 2020. Computer-aided classification of suspicious pigmented lesions using wide-field images. Computer Methods and Programs in Biomedicine 195, 105631. doi:https://doi.org/10.1016/j.cmpb.2020.105631.
- [6] Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K., 2019. This looks like that: Deep learning for interpretable image recognition. Advances in Neural Information Processing Systems 32, 8930–8941.
- [7] Chen, Z., Bei, Y., Rudin, C., 2020. Concept whitening for interpretable image recognition. Nature Machine Intelligence 2, 772–782.
- [8] Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al., 2018. 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. pp. 168–172.
- [9] Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al., 2019. Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288.
- [10] Coppola, D., Lee, H.K., Guan, C., 2020. Interpreting mechanisms of prediction for skin cancer diagnosis using multi-task learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 734–735.
- [11] Council of the European Union, 2016. 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.
- [12] Data Language (UK) Ltd, 2021. Data [29 language's explainable ai platform. https://datalanguage.com/products/datalanguageai/explainable-ai-platform. Accessed: 2021-08-15.
- [13] Decoded Health, 2021. The world's first clinical hyperautomation platform - a force multiplier for physicians. https://www.decodedhealth.com/. Accessed: 2021-08-15.

- [14] de Faria, S.M., Filipe, J.N., Pereira, P.M., Tavora, L.M., Assuncao, P.A., Santos, M.O., Fonseca-Pinto, R., Santiago, F., Dominguez, V., Henrique, M., 2019. Light field image dataset of skin lesions, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE. pp. 3905–3908.
- [15] Fong, R., Patrick, M., Vedaldi, A., 2019. Understanding deep networks via extremal perturbations and smooth masks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2950–2958.
- [16] Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D.I., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., et al., 2018. Niftynet: a deep-learning platform for medical imaging. Computer Methods and Programs in Biomedicine 158, 113–122.
- [17] Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S., 2020. Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE transactions on medical imaging 40, 699– 711.
- [18] Hacarus Inc., 2021. Hacarus sparse modeling based ai, edge ai with learning and inference capability, white box ai. https://hacarus.com/. Accessed: 2021-08-15.
- [19] Hasan, M.K., Roy, S., Mondal, C., Alam, M.A., Elahi, M.T.E., Dutta, A., Raju, S.T.U., Jawad, M.T., Ahmad, M., 2021. Dermo-doctor: A framework for concurrent skin lesion detection and recognition using a deep convolutional neural network with end-to-end dual encoders. Biomedical Signal Processing and Control 68, 102661.
- [20] Hendricks, L.A., Hu, R., Darrell, T., Akata, Z., 2018. Grounding visual explanations, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 264–279.
- [21] Izadyyazdanabadi, M., Belykh, E., Cavallo, C., Zhao, X., Gandhi, S., Moreira, L.B., Eschbacher, J., Nakaji, P., Preul, M.C., Yang, Y., 2018. Weakly-supervised learning-based feature localization for confocal laser endomicroscopy glioma images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 300–308.
- [22] Jiang, S., Li, H., Jin, Z., 2021. A visually interpretable deep learning framework for histopathological image-based skin cancer diagnosis. IEEE Journal of Biomedical and Health Informatics 25, 1483–1494.
- [23] Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G., 2019. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE Journal of Biomedical and Health Informatics 23, 538–546. doi:10.1109/JBHI.2018.2824327.
- [24] Khan, A.I., Shah, J.L., Bhat, M.M., 2020. Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images. Computer Methods and Programs in Biomedicine 196, 105581.
- [25] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al., 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV), in: International conference on machine learning, PMLR. pp. 2668–2677.
- [26] Kittler, H., Marghoob, A.A., Argenziano, G., Carrera, C., Curiel-Lewandrowski, C., Hofmann-Wellenhof, R., Malvehy, J., Menzies, S., Puig, S., Rabinovitz, H., et al., 2016. 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 74, 1093–1106.
- [27] Lucieri, A., Bajwa, M.N., Braun, S.A., Malik, M.I., Dengel, A., Ahmed, S., 2020a. On interpretability of deep learning based skin lesion classifiers using concept activation vectors, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE. pp. 1–10.
- [28] Lucieri, A., Bajwa, M.N., Dengel, A., Ahmed, S., 2020b. Achievements and challenges in explaining deep learning based computeraided diagnosis systems. arXiv preprint arXiv:2011.13169.
- [29] Lucieri, A., Bajwa, M.N., Dengel, A., Ahmed, S., 2020c. Explaining AI-based Decision Support Systems using Concept Localization Maps, in: International Conference on Neural Information Processing, Springer. pp. 185–193.
- [30] Lucieri, A., Dengel, A., Ahmed, S., 2021. Deep learning based decision support for medicine–a case study on skin cancer diagnosis.

arXiv preprint arXiv:2103.05112.

- [31] Maglogiannis, I., Delibasis, K.K., 2015. Enhancing classification accuracy utilizing globules and dots features in digital dermoscopy. Computer Methods and Programs in Biomedicine 118, 124–133.
- [32] Mahbod, A., Tschandl, P., Langs, G., Ecker, R., Ellinger, I., 2020. The effects of skin lesion segmentation on the performance of dermatoscopic image classification. Computer Methods and Programs in Biomedicine 197, 105725.
- [33] Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R., Rozeira, J., 2013. 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. pp. 5437–5440.
- [34] Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences 116, 22071– 22080.
- [35] Nachbar, F., Stolz, W., Merkle, T., Cognetta, A.B., Vogt, T., Landthaler, M., Bilek, P., Braun-Falco, O., Plewig, G., 1994. The abcd rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. Journal of the American Academy of Dermatology 30, 551–559.
- [36] Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144.
- [37] Society, A.C., 2020. Cancer facts & figures 2020. Am. Cancer Soc .
- [38] Sonntag, D., Nunnari, F., Profitlich, H.J., 2020. The skincare project, an interactive deep learning system for differential diagnosis of malignant skin lesions. arXiv preprint arXiv:2005.09448.
- [39] Sunija, A., Kar, S., Gayathri, S., Gopi, V.P., Palanisamy, P., 2021. Octnet: A lightweight cnn for retinal disease classification from optical coherence tomography images. Computer Methods and Programs in Biomedicine 200, 105877.
- [40] Tosun, A.B., Pullara, F., Becich, M.J., Taylor, D.L., Chennubhotla, S.C., Fine, J.L., 2020. Histomapr[™]: An explainable ai (xai) platform for computational pathology solutions, in: Artificial Intelligence and Machine Learning for Digital Pathology. Springer, pp. 204–227.
- [41] Tschandl, P., Rosendahl, C., Akay, B.N., Argenziano, G., Blum, A., Braun, R.P., Cabo, H., Gourhant, J.Y., Kreusch, J., Lallas, A., et al., 2019. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. JAMA dermatology 155, 58– 65.
- [42] Tschandl, P., Rosendahl, C., Kittler, H., 2018. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data 5, 1–9.
- [43] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., Hu, X., 2020. Score-cam: Score-weighted visual explanations for convolutional neural networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 24–25.
- [44] Xiang, A., Wang, F., 2019. Towards interpretable skin lesion classification with deep learning models, in: AMIA annual symposium proceedings, American Medical Informatics Association. p. 1246.
- [45] Young, K., Booth, G., Simpson, B., Dutton, R., Shrapnel, S., 2019. Deep neural network or dermatologist?, in: Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support. Springer, pp. 48–55.
- [46] Zhang, Z., Xie, Y., Xing, F., McGough, M., Yang, L., 2017. Mdnet: A semantically and visually interpretable medical image diagnosis network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6428–6436.
- [47] Zicari, R.V., Ahmed, S., Amann, J., Braun, S.A., Brodersen, J., Bruneault, F., Brusseau, J., Campano, E., Coffee, M., Dengel, A., et al., 2021. Co-design of a trustworthy ai system in healthcare: Deep learning based skin lesion classifier. Frontiers in Human Dynamics , 40.



Adriano Lucieri completed his BE in Mechatronic Engineering from Duale Hochschule Baden-Württemberg (DHBW) Mannheim and MS in Mechatronic Systems Engineering from Hochschule Pforzheim in Germany. He is presently pursuing PhD from Technische Universität Kaiserslautern (TUK), Germany and is also working as Research Assistant at Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI). His research focus lies on improving the explainability and transparency of Computer-Aided Diagnosis (CAD) systems based on Deep Learning for medical image analysis. His work includes conceptbased explanation of skin lesion classifiers as well as the localisation of concept regions in input images.



Andreas Dengel is Scientific Director at DFKI GmbH in Kaiserslautern. In 1993, he became Professor in Computer Science at TUK where he holds the chair Knowledge-Based Systems. Since 2009 he is appointed Professor (Kyakuin) in Department of Computer Science and Information Systems at Osaka Prefecture University. He received his Diploma in CS from TUK and his PhD from University of Stuttgart. He also worked at IBM, Siemens, and Xerox Parc. Andreas is member of several international advisory boards, has chaired major international conferences, and founded several successful start-up companies. He is co-editor of international computer science journals and has written or edited 12 books. He is author of more than 300 peer-reviewed scientific publications and supervised more than 170 PhD and master theses. Andreas is an IAPR Fellow and received many prominent international awards. His main scientific emphasis is in the areas of Pattern Recognition, Document Understanding, Information Retrieval, Multimedia Mining, Semantic Technologies, and Social Media.

40 publications including several journal and high

ranked conference papers.



Sheraz Ahmed is Senior Researcher at DFKI GmbH in Kaiserslautern, where he is leading the area of Time Series Analysis. He received his MS and PhD degrees in Computer Science from TUK, Germany under the supervision of Prof. Dr. Prof. h.c. Andreas Dengel and Prof. Dr. habil. Marcus Liwicki. His PhD topic is Generic Methods for Information Segmentation in Document Images. Over the last few years, he has primarily worked on development of various systems for information segmentation in document images. His research interests include document understanding, generic segmentation framework for documents, gesture recognition, pattern recognition, data mining, anomaly detection, and natural language processing. He has more than 30 publications on the said and related topics including three journal papers and two book chapters. He is a frequent reviewer of various journals and conferences including Patter Recognition Letters, Neural Computing and Applications, IJDAR, ICDAR, ICFHR, and DAS.



Muhammad Naseer Bajwa completed his BS in Computer Engineering from COMSATS Institute of Information Technology (CIIT), Pakistan and MS in Computer Engineering from King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia. He is presently pursuing PhD from Technische Universität Kaiserslautern (TUK), Germany and is also working as Research Assistant at Deutschen Forschungszentrums für Künstliche Intelligenz GmbH (DFKI). His main area of research is towards realising a practically usable, confident and interpretable Computer-Aided Diagnosis (CAD) system. He has published his works in various peer reviewed journals and top ranked conferences on detection of ocular disorders like glaucoma and diabetic retinopathy using retinal fundus images, automated diagnosis of cutaneous diseases using dermoscopic images, curation of retinal fundus images dataset for glaucoma detection and segmentation of optic disc and cup (G1020), and interpretability of CAD for skin lesions.



Stephan Alexander Braun is a board-certified dermatologist and dermatopathologist and works at the University Hospital of Münster and Düsseldorf, Germany. His scientific work focuses on the diagnosis and treatment of skin tumors.



Muhammad Imran Malik received his master's and PhD degrees in Artificial Intelligence, in 2011 and 2015 respectively, from the University of Kaiserslautern. He also worked in the German Research Center for Artificial Intelligence GmbH (DFKI), Kaiserslautern, Germany. His Ph.D. topic was automated forensic handwriting analysis on which he focused on both the perspectives of forensic handwriting examiners and pattern recognition researchers. He is currently an Assistant Professor with the School of Electrical Engineering and Computer Science (SEECS) at the National University of Sciences and Technology (NUST), Islamabad, Pakistan. He has authored more than