Computer Methods and Programs in Biomedicine 215 (2022) 106620

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

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], Andreas Dengel[a,b], Sheraz Ahmed[a]

[a] German Research Center for Artificial Intelligence (DFKI) GmbH, Trippstadter Straße 122, 67663 Kaiserslautern, Germany
[b] Technical University Kaiserslautern, Erwin-Schrödinger-Straße 52, 67663 Kaiserslautern, Germany
[c] University Hospital Münster, Albert-Schweitzer-Campus 1, 48149 Münster, Germany
[d] University Hospital of Düsseldorf, Moorenstraße 5, 40225 Düsseldorf, Germany
[e] School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan
[f] Deep Learning Laboratory, National Center of Artificial Intelligence, Islamabad, Pakistan

ARTICLE INFO

Article history:
Received 25 August 2021
Revised 1 December 2021
Accepted 3 January 2022

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 the successful deployment of Artificial Intelligence (AI) based Computer-Aided Diagnosis (CAD) systems in everyday clinical workflows is their lack of transparent decision-making. Although commonly used eXplainable AI (XAI) methods provide insights into these largely opaque algorithms, such explanations are usually convoluted and not readily comprehensible. The explanation of decisions regarding the malignancy of skin lesions from dermoscopic images demands particular clarity, as the underlying medical problem definition is ambiguous in itself. This work presents ExAID (Explainable AI for Dermatology), a novel XAI framework for biomedical image analysis that provides multi-modal concept-based explanations, consisting of easy-to-understand textual explanations and visual maps, to justify the predictions.

Methods: Our framework relies on Concept Activation Vectors to map human-understandable concepts to those learned by an arbitrary Deep Learning (DL) based algorithm, and Concept Localisation Maps to highlight those 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 presented in a diagnostic interface for use in clinical routines. Moreover, the framework includes an educational mode providing dataset-level explanation statistics as well as 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 publicly available dermoscopic image datasets, we show the utility of multi-modal explanations for CAD-assisted scenarios even in case of wrong disease predictions. We demonstrate that concept detectors for the explanation of pre-trained networks reach accuracies of up to 81.46%, which is comparable to supervised networks trained end-to-end.

Conclusions: We present a new end-to-end framework for the multi-modal explanation of DL-based biomedical image analysis in Melanoma classification and evaluate its utility on an array of datasets. Since perspicuous explanation is one of the cornerstones of any CAD system, we believe that ExAID will accelerate the transition from AI research to practice by providing dermatologists and researchers with an effective tool that they can both understand and trust. ExAID can also serve as the basis for similar applications in other biomedical fields.

* Corresponding author at: German Research Center for Artificial Intelligence (DFKI) GmbH, Trippstadter Straße 122, 67663 Kaiserslautern, Germany.
E-mail 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).

https://doi.org/10.1016/j.cmpb.2022.106620
0169-2607/

1. Introduction

In 2016, Ribeiro et al. [1] reported an image classifier that was able to inadvertently classify correctly but for the wrong reasons. They found that their wolf versus dog classifier learnt an undesirable correlation between the wolf class and snowy background. The classifier would predict the presence of a wolf in an image if there was snow in the background. If it were not due to the authors' vigilance in finding explanations to the model's predictions, it would have been difficult to thoroughly evaluate the 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-critical domains resulting from such misunderstandings can potentially have a grave impact on human lives. Therefore, despite ubiquitous utilisation of Deep Learning (DL) methods for Computer-Aided Diagnosis (CAD) in the last decade [2–4], the hesitation of medical practitioners in trusting diagnostic predictions of such systems is understandable since they often provide little to no cognisance regarding their decision-making process [5]. In addition to evaluating the reasons behind a model's predictions, explanations can also help in revealing new diagnostic criteria [6] previously unknown to medical practitioners. The requirement for a CAD to be explainable arose with early applications of Artificial Intelligence (AI) in healthcare. However, it became more relevant with recent ethical and legal standards [7,8].

Explainable AI (XAI) methods for image-based classifiers come in a variety of forms and provide explanations using different modalities such as feature-relevance visualisations [1,9–11], textual explanations [12,13], or quantitative relevance measures for abstract concepts [14,15]. They differ not only in the way they are presented to their users but also in their derivation, resulting in varying levels of insight regarding the decision-making of the AI. Furthermore, these methods can be either ante-hoc (e.g. ProtoPNet [9], MDNet [13]), 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 [11], TCAV [15]) or model-agnostic (e.g. LIME [1], EP [10]) techniques [16]. 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 [17]. 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 ABCD-rule [18] or 7-point checklist [19] to judge the malignancy of a lesion. Currently, AI-based dermatology focuses mostly on the analysis of dermoscopic images [20–22]. However, first approaches towards analysing raw, clinical images have been proposed as well [23]. The majority of explanation approaches for dermoscopic skin lesion analysis rely on the application of visual XAI through saliency maps [24,25] or attention mechanisms [26,27]. The direct detection and localisation of dermoscopic criteria as used by doctors in manual classification provides more intelligibility by design. Coppola et al. [28], for instance, trained a multi-task Convolutional Neural Network (CNN) predicting dermoscopic features with information sharing between different subnetworks to increase interpretability. In [29], Lucieri et al. applied the concept-based TCAV method predicting dermoscopic criteria from pre-trained network embeddings, to explain the network's predictions. Dermoscopic criteria localisation has been achieved by combining perturbation-based saliency methods with TCAV in Lucieri et al. [30] and through explicit segmentation of criteria in Kawahara and Hamarneh [31], Sonntag et al. [32]. For a complete survey on XAI in dermatology, the reader is referred to [33].

Several frameworks for AI-based medical image classification have been proposed in recent years [32,34–36]. 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 have also emerged [37–40], 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 provides an end-to-end explanation pipeline for arbitrary DL-based skin lesion models, while being adaptable to any other biomedical imaging use case. Instead of relying on heuristics and auxiliary models for its explanations, ExAID provides easy-to-understand multi-modal explanations, which directly depend on the DL classifier, and centre around common dermoscopic biomarkers. With two separate interfaces, our framework targets both practising dermatologists and researchers from medicine and computer science. The integrated and human-centric approach of ExAID acts as a bridge between DL researchers and medical professionals, bringing both domains closer together and accelerating the transition of state-of-the-art DL into diagnostic processes and medical research.

The contributions of our work are threefold. First, we introduce a new concept-based textual explanation method with adaptive granularity and highlighting of contraindications. Second, we integrate different human-aligned explanation modalities (quantitative, visual, and textual) within an intuitive framework for intelligible CAD, providing different modes for clinical use and in-depth analysis of model and data. Lastly, we provide an extensive evaluation of our framework and the utility of its explanations on the example of skin lesion classification from dermoscopic images.

2. Methods

2.1. Datasets

ExAID contains two types of classifiers: Concept-level classifiers for the detection of dermatological concepts in a given image and the disease-level classifier for lesion diagnosis that will be explained. To train these two classifiers, datasets with two types of labels are required, namely concept annotations (presence or absence of dermoscopic concepts) and disease labels (*Melanoma* and *Nevus*).

2.1.1. Datasets for concept-level classification

Training of concept classifiers requires annotations regarding the 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 for concept detection primarily to PH2 [41] and Derm7pt [42]. 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 well-curated annotations regarding the presence or absence of various concepts. The Derm7pt dataset contains 1011 clinical and dermoscopic images classified into one of four diagnosis classes or a miscellaneous class. Only 823 images from this dataset, belonging to *Melanoma* and *Nevi* classes, have been considered. The combination of Derm7pt and PH2 used for concept classification is subsequently referred to as D7PH2. Table 1

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

Table 1
Distribution of data in training, validation, and test splits for concept-level classification with D7PH2 dataset.

Split	Dataset	Lesions		
		Melanoma	Nevi	Total
Train	Derm7pt	158	368	526
	PH2	26	102	128
Validate	Derm7pt	40	92	132
	PH2	6	26	32
Test	Derm7pt	50	115	165
	PH2	8	32	40
Total	Derm7pt	248	575	823
	PH2	40	160	200

Table 2
Distribution of data in training, validation, and test splits for disease-level classification.

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

shows the distribution of images used in the concept-level classification task. The ISIC[2] 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.

2.1.2. Datasets for disease-level classification

The training set for disease-level classification consists of *Melanoma* and *Nevi* images taken from ISIC 2019, PH2 and Derm7pt datasets. ISIC 2019 challenge dataset is a public collection of 25,331 images of different provenance divided into eight different classes. Since the common denominator of all three datasets are *Melanoma* and *Nevi* classes, we assembled a subset consisting of images from these two classes only, and manually cleaned the dataset for duplicates and samples with low quality (e.g. systematic artefacts), resulting in a total of 6475 images. As PH2 and Derm7pt are also used for training the concept-level classifiers, a custom dataset split is obtained by combining 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 2. The generalisability of the model is evaluated on some other datasets including the 2016 and 2017 ISIC challenge datasets and the SKINL2 [43] dataset.

2.2. ExAID framework

At its core, ExAID is a generic toolbox for human-centred post-hoc explanations and is 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 dermoscopic concepts to their corresponding representation learnt by the DL-based model using the Testing with Concept Activation Vector (TCAV) method [15]. A linear binary classifier is trained for the detection of a dermatological concept C from the model's intermediate activations $a_l(x)$ at the lth layer. The corresponding Concept Activation Vector (CAV) $\vec{v}_C$ is the vector normal to the resulting hyperplane and represents the main concept direction as expressed in the latent space. The TCAV score is a quantitative explanation metric which estimates the global influence of a concept C on the classifier when making predictions for a specific class label k. The metric is defined as:

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

where X_k denotes all inputs labelled as k and $S_{C,k,l}(a_l(x))$ being the directional derivative of a sample's activation a from layer l with respect to class k and concept C.

Once all CAVs are trained on PH2 and Derm7pt datasets, the Concept Identifier is able to predict the presence or absence of individual concepts on unseen images based only on the model's latent activations. Furthermore, it provides TCAV scores estimating a concept's overall contribution in predicting a target class. CAVs can be computed on any intermediate model layer. Our framework allows the inspection of CAVs from arbitrary layers for detailed investigation but also includes an automatic selection strategy for clinical diagnosis settings. In this case, layers and corresponding CAVs for a concept are selected based on the best accuracy obtained on the concept-level validation datasets. More details on the Concept Identifier and the CAV training procedure can be found in Lucieri et al. [29].

2.2.2. Concept localiser

Concept Localisation Maps (CLMs) [30] extend CAVs by localising regions pertinent to a learnt concept in the latent space of a trained image classifier. They provide qualitative and quantitative assurance of the model's ability to learn the correct interpretation of a concept by indicating the exact spatial location that contributed to concept prediction and 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 Lucieri et al. [30] to generate spatial importance values based on variation of 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 through 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 rule-base 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 on the given

[2] ISIC datasets available under https://www.isic-archive.com.

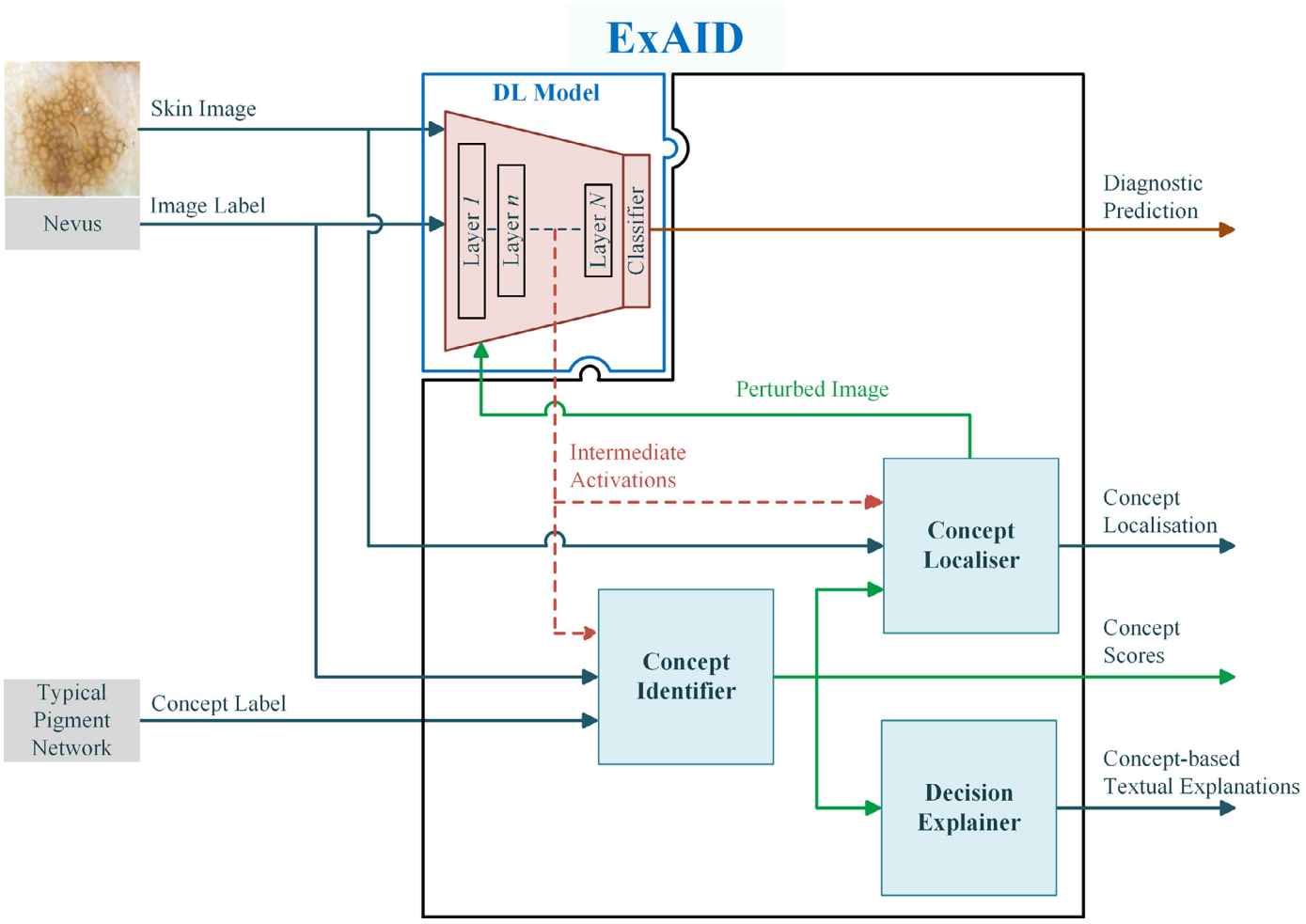

Fig. 1. ExAID Framework architecture. The schematic drawing shows the input, output, and flow of information through ExAID as well as the relationship between its components.

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 of coherent and easy-to-understand explanation texts. An explanation sentence is constructed based on concept predictions and directional derivatives computed during concept detection under discrimination between absence, moderate evidence, and strong evidence of concepts to reflect the fuzzy nature of the concepts' appearances. Manifestation of a concept is decided using 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 and True Positive Rate among all positive predictions on the concept training dataset for moderate and strong evidence thresholds, respectively. The directional derivatives of the predicted class along the individual CAV are used to indicate a positive or negative influence of concept on the prediction. Conceptual evidence is listed after the keyword "despite" in case of negative class influence to signalise contraindication.

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 a 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 the visual examination of the patient's skin lesions to decide on further investigation of a potentially malignant lesion. Provided enough evidence for malignancy, the suspicious tissue may be excised under local anaesthesia. Physicians with considerable experience in dermoscopy develop an intuition that allows them to promptly conclude while novices initially need to pay greater attention to the assessment of a particular skin lesion. This is, among other things, owing 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 that stimulates the physician's thought process and breaks the routine. By providing explicit explanations 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. Additional CLM visualisations can be individually acti-

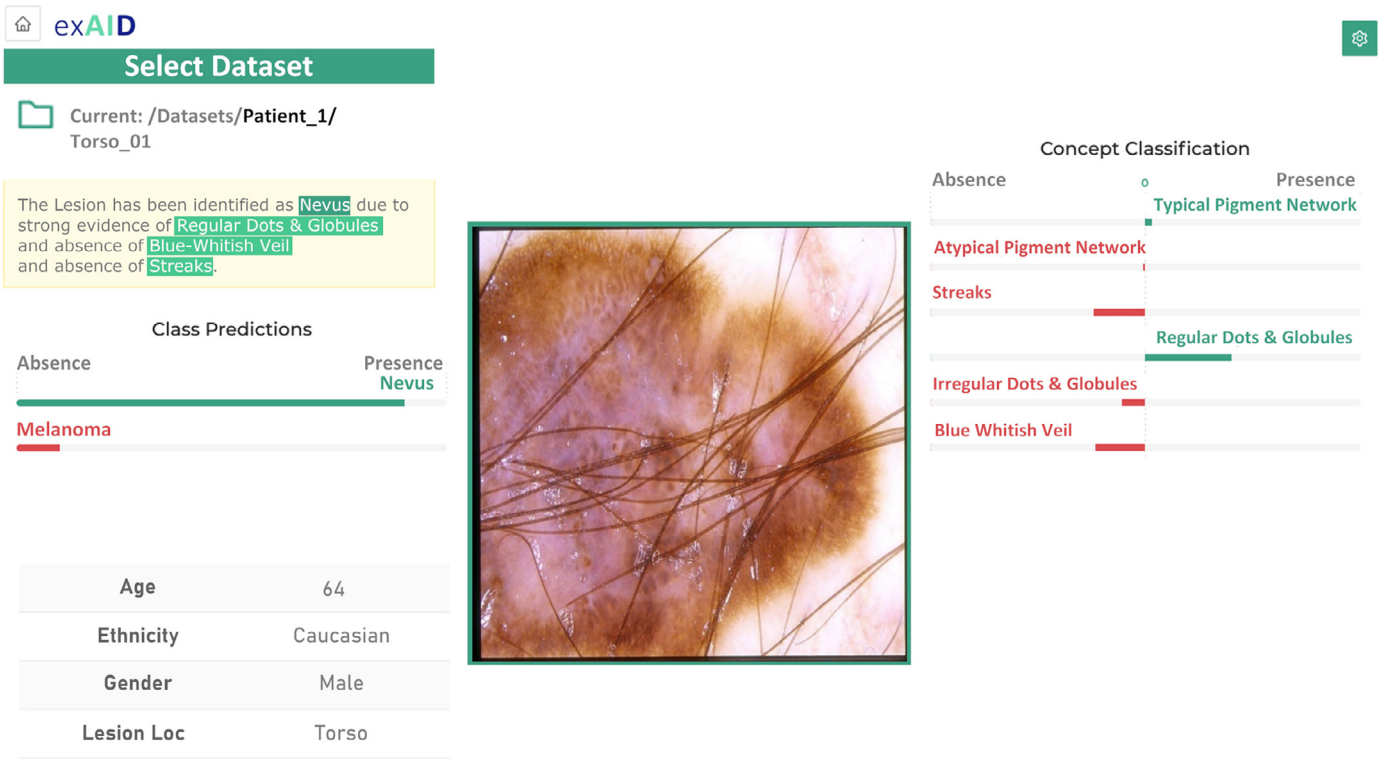

Fig. 2. The Diagnostic Mode of ExAID is intended as a Decision Support System that integrates into routine clinical workflows. The standard model prediction is augmented by quantitative, visual, and textual concept explanations. Visual CLMs can be toggled by clicking on the different concept classifications.

vated by clicking on concept scores. While allowing physicians to examine the dermoscopic image manually, an initial diagnosis suggestion is provided, supported by concept-based 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-making process by considering and validating biomarker scores along with their optional localisations provided in the form of visual CLMs.

2.3.2. Educational mode

The explanation of a classifier's 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, the 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 filtering 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 spotlighting certain useful properties of samples to further facilitate the 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 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 the 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 projector[3] 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 combinations,[4] 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

[3] https://projector.tensorflow.org/.

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

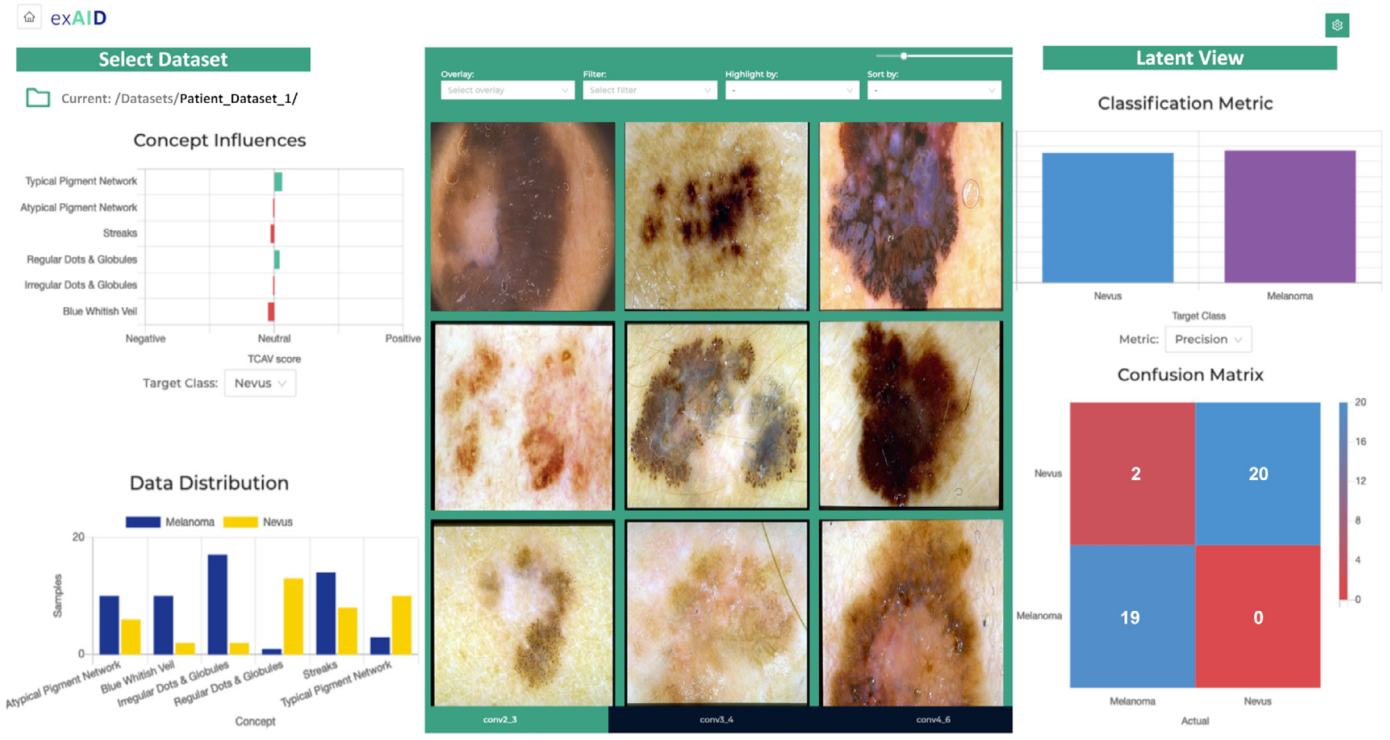

Fig. 3. The Educational Mode of ExAID is intended for use in medical research and education. Its interface allows the exploration of data composition and model behaviour through a row of interactive features, including data filtering, highlighting, statistics computation, as well as the visualisation of model performance and explanation metrics such as the dataset-wide TCAV score.

Table 3
Performance evaluation of lesion classifier on various datasets.

Datasets	*N*	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

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 Lucieri et al. [29] is followed. 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 labels. 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 based on 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 the evaluation of only two concepts each. Table 4 presents Macro Average F1-Scores for concept detection.

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	79.66	63.55	61.94	71.18
ISIC2017	51.75	50.86	–	–	–
ISIC2016	56.53	–	53.03	–	–

Table 5
Performance evaluation of concept classifiers compared with results from Kawahara et al. [42]. Values are given as accuracies.

Method	Streaks	Pigment Netw.	Dots & Glob.	Regr. Struct.	Blue-Whit. Veils
Ours	73.66	81.46	70.73	64.88	74.63
Kawahara et al.	74.20	70.9	60.00	77.20	87.10

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 79.66%. 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 50.86%, respectively.

Table 5 compares the concept detectors' test accuracies with the results of supervised dermoscopic feature classifiers reported in Kawahara et al. [42]. It is evident that our concept detectors perform comparable even though we do not fine-tune the feature extractor on the task.

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 binarisation scheme, subjectivity of concept annotations as well as

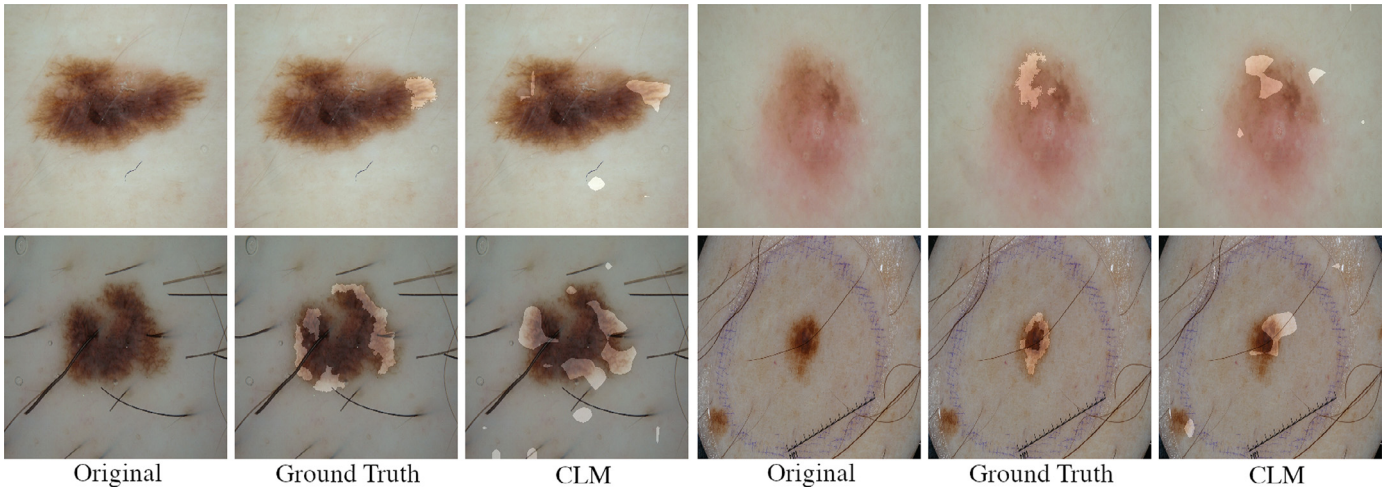

Fig. 4. Positive and negative examples of visual explanations provided by ExAID, along with the corresponding samples and ground truth concept masks.

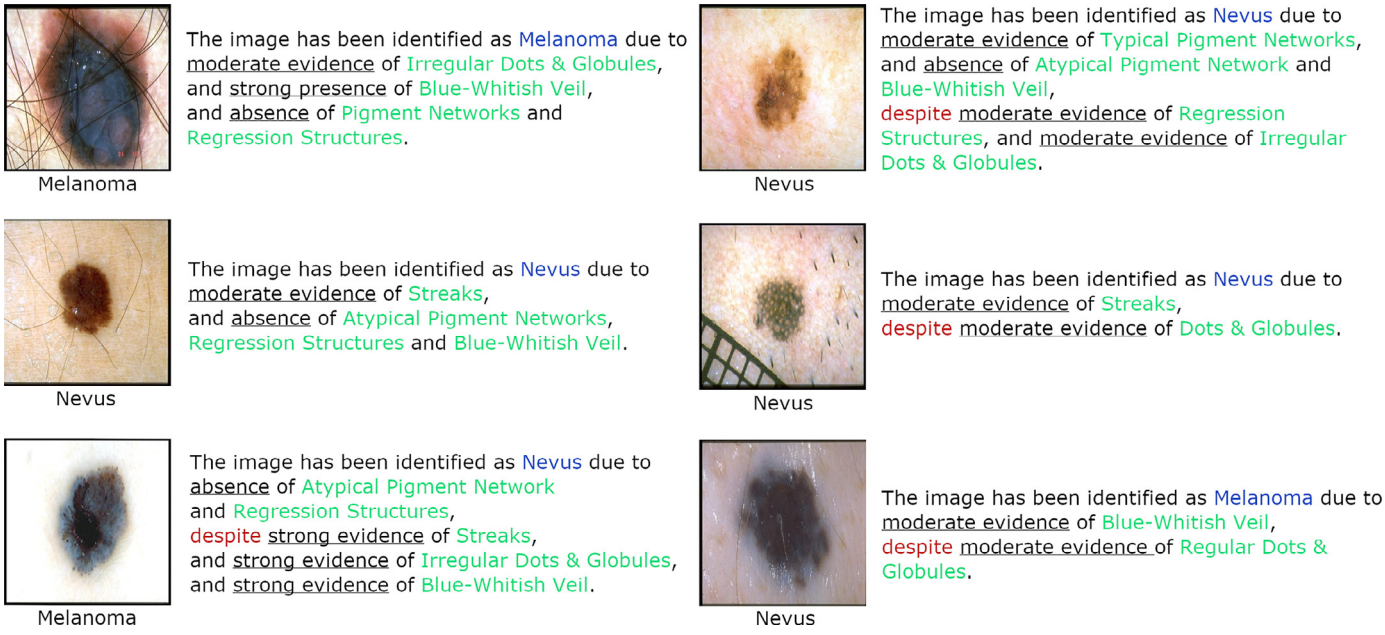

Fig. 5. Positive and negative examples of textual explanations provided by ExAID along with the corresponding skin lesion samples. The ground truth class of the sample is given below the image.

lack of representative metrics for fuzzy localisation tasks. Proper binarisation is especially 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. Fig. 4 shows examples of the model's concept localisation ability for classes *Streaks* and *Pigment Network* using an adaptation of the method proposed in Lucieri et al. [30]. 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. Fig. 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 correct explanations for a classifier's decisions. In the following, previous insights are discussed and qualitative results are analysed, followed by a comparison to existing frameworks and a detailed discussion about the limitations.

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 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 Fig. 4. The 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 highlight noise problems inherent in perturbation-based CLM computation and the dependence on a proper binarisation scheme.

4.4. Textual explanation

The examples in Fig. 5a depict instances with correct concept predictions, showcasing the simplicity and intelligibility of the generated explanations. An explanation text briefly reflects the most important criteria necessary for experts to understand the network's decision. Interestingly, it appeared that although correct concept predictions were given, the network sometimes misclassified the underlying disease as seen in the third row of Fig. 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.

Fig. 5 b 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 as well as contraindications more frequently as compared to the samples from Fig. 5a. This clearly shows that irrespective of the model used for prediction, ExAID can provide well-founded justifications which help to express model uncertainty, encouraging closer examination of rare and edge cases.

4.5. Comparison to other frameworks

ExAID is a multi-purpose framework for DL-based medical image analysis with intuitive interfaces for clinical and educational settings. Most published works develop their frameworks in the form of research code or low-level tools for developing or providing medical imaging pipelines. A majority of tools such as NiftyNet [34], DRANet [36], and Dermo-DOCTOR [35] are targeted towards computer scientists and AI developers, providing very simplistic, or no interfaces at all. This makes such work inaccessible for experts trained solely in medical domains, limiting its practical impact. In addition, those methods usually do not include XAI aspects or limit such components to the visualisation of basic attribution or attention maps (e.g. DRANet).

The tools which relate most to our proposed work are the Skincare project [32] and HistoMapr™ [40]. In [32], an interactive DL-based decision support system for dermatology is proposed. The Skincare framework consists of several DL components like VGG16 classifiers for binary and 8-class classification of lesions, several UNet-based architectures for lesion and feature segmentation, as well as further handcrafted features measuring asymmetry and colour aspects, among others. For the explanation of their DL-based classifiers, attribution methods like GradCAM and Randomized Input Sampling (RISE) are utilised. The results of segmentation and handcrafted computation of features convey the impression of an additional justification of the AI diagnosis. However, due to the separate nature of their computations, they can only serve as an additional stimulus for the medical practitioner's own reasoning. A lack of understanding of the underlying technical details might lead to misinterpretation of these results by end-users. Moreover, handcrafted explanations are not informative for model analysis and hence invaluable for computer scientists.

HistoMapr™ is a commercial Machine Learning framework for the support of experts during pathological investigations. As compared to most AI-based frameworks, HistoMapr™ emphasises user-centricity by providing explanations in domain-specific terminology. Being a combination of statistical and Machine Learning methods, the framework provides different hints to support medical professionals and streamline their decision processes. However, similar to the Skincare Project, their explanation components do not reveal details about the actual decision process of a deep model.

In contrast to previous AI-based frameworks, ExAID follows an approach that is both user-centric as well as end-to-end DL-based. The explanations provided by our framework depend solely on the classification model at hand, as well as a small amount of expert-curated concept data. The utilisation of concept-based explanation approaches for DL guarantees human intelligibility without the cost of unfaithful explanations from unrelated modules. Moreover, to the best of our knowledge, ExAID is the first medical framework providing coherent, multi-modal concept-based explanations consisting of quantitative, visual, and textual components. The flexible interface for the explanation of an arbitrary end-to-end DL classifier, with arbitrary, domain-specific concepts makes the framework easily adaptable to other medical tasks. With its two interfaces, ExAID targets not only medical domain experts but also computer scientists for the analysis of the DL classifier.

4.6. 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 that holistically includes a wide variety of stakeholders [44]. This as-

sures not only practical value, but also compliance with ethical aspects from an early stage. 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 a low sample quality attributable to a lack of process standardisation,[5] 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 a 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 that results in noisy heatmaps and high sensitivity to hyperparameters, especially in the case of varying biomarker sizes. Future work applying optimisation-based perturbation methods for concept localisation will mitigate those issues, resulting in more 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. Incorporating 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 the 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 are required, which will reflect higher-quality explanations. Moreover, evaluation of CLMs is aggravated by noise artefacts emerging during binarisation and a lack of definite measures for fuzzy localisation tasks. A qualitative evaluation in a 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 is not only reflected in the data annotations, but also in the general uncertainty surrounding the field of dermoscopy. Despite first attempts towards the standardisation of dermoscopic terminologies and concepts [45], no 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 [19] as well as further dermoscopic concepts from Mendonça et al. [41], due to the public availability of annotated data. The commitment to a specific set of concepts before 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. Before 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] Different camera setups, operators, and techniques like polarised and non-polarised dermoscopy resulting in varying image quality, lighting, alignment, and artefacts.

5. Conclusion

With ExAID, this article presents a framework that consolidates our previous works on concept-based explanations for skin lesion diagnosis, supplemented by novel concept-based textual explanations to offer coherent and intuitive justifications of the models' predictions. We showed that our framework can generate meaningful explanations while requiring only a small amount of concept-annotated data. The analysis showed that the pre-trained network can better detect and locate coarse-grained biomarkers like *Streaks, Pigment Networks* and *Blue-Whitish Veils* for explanations, as compared to fine-grained ones. We proposed a new method for concept-based textual explanations with graded concept manifestations and the option to highlight contraindications, demonstrating the method's utility not only in the case of correct classifications but especially when a lesion is misclassified.

Our presented framework offers human-aligned and multimodal explanations with intelligible interfaces targeting both medical practitioners as well as researchers. As compared to other frameworks, it focuses on the faithful explanation of the underlying decision process instead of constructing unrelated explanations and adapts to other biomedical image tasks. We showed that, despite the limitation in terms of data and annotation availability, the system already provides first useful insights into a DL classifier's decision-making process, even in the case of a wrong prediction. Furthermore, we identified potential improvements in concept localisation and concept selection. 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 biomedical domains like radiology or histology.

Statement of ethical approval

Ethical approval is not required for retrospective studies, not reporting on primary research. All data analysed is publicly available.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

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

[1] M.T. Ribeiro, S. Singh, C. Guestrin, "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, 2016, pp. 1135–1144.
[2] M.A. Al-Antari, S.-M. Han, T.-S. Kim, Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms, Comput. Methods Prog. Biomed. 196 (2020) 105584.
[3] A.I. Khan, J.L. Shah, M.M. Bhat, Coronet: a deep neural network for detection and diagnosis of COVID-19 from chest X-ray images, Comput. Methods Prog. Biomed. 196 (2020) 105581.
[4] A.P. Sunija, S. Kar, S. Gayathri, V.P. Gopi, P. Palanisamy, Octnet: a lightweight CNN for retinal disease classification from optical coherence tomography images, Comput. Methods Prog. Biomed. 200 (2021) 105877.
[5] A. Lucieri, M.N. Bajwa, A. Dengel, S. Ahmed, Achievements and challenges in explaining deep learning based computer-aided diagnosis systems, arXiv preprint arXiv:2011.13169 (2020).

[6] M. Izadyyazdanabadi, E. Belykh, C. Cavallo, X. Zhao, S. Gandhi, L.B. Moreira, J. Eschbacher, P. Nakaji, M.C. Preul, Y. Yang, Weakly-supervised learning-based feature localization for confocal laser endomicroscopy glioma images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 300–308.
[7] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[8] 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), 2016, (available at http://data.europa.eu/eli/reg/2016/679/2016-05-04).
[9] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, J.K. Su, This looks like that: deep learning for interpretable image recognition, Adv. Neural Inf. Process. Syst. 32 (2019) 8930–8941.
[10] R. Fong, M. Patrick, A. Vedaldi, Understanding deep networks via extremal perturbations and smooth masks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2950–2958.
[11] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, X. Hu, Score–CAM: score-weighted visual explanations for convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 24–25.
[12] L.A. Hendricks, R. Hu, T. Darrell, Z. Akata, Grounding visual explanations, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 264–279.
[13] Z. Zhang, Y. Xie, F. Xing, M. McGough, 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.
[14] Z. Chen, Y. Bei, C. Rudin, Concept whitening for interpretable image recognition, Nat. Mach. Intell. 2 (12) (2020) 772–782.
[15] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al., Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV), in: International Conference on Machine Learning, PMLR, 2018, pp. 2668–2677.
[16] W.J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu, Definitions, methods, and applications in interpretable machine learning, Proc. Natl. Acad. Sci. 116 (44) (2019) 22071–22080.
[17] A.C. Society, Cancer facts & figures 2020, Am. Cancer Soc. (2020) https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf.
[18] F. Nachbar, W. Stolz, T. Merkle, A.B. Cognetta, T. Vogt, M. Landthaler, P. Bilek, O. Braun-Falco, G. Plewig, The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions, J. Am. Acad. Dermatol. 30 (4) (1994) 551–559.
[19] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, M. Delfino, 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, Arch. Dermatol. 134 (12) (1998) 1563–1570.
[20] M.N. Bajwa, K. Muta, M.I. Malik, S.A. Siddiqui, S.A. Braun, B. Homey, A. Dengel, S. Ahmed, Computer-aided diagnosis of skin diseases using deep neural networks, Appl. Sci. 10 (7) (2020) 2488.
[21] A. Mahbod, P. Tschandl, G. Langs, R. Ecker, I. Ellinger, The effects of skin lesion segmentation on the performance of dermatoscopic image classification, Comput. Methods Prog. Biomed. 197 (2020) 105725.
[22] P. Tschandl, C. Rosendahl, B.N. Akay, G. Argenziano, A. Blum, R.P. Braun, H. Cabo, J.-Y. Gourhant, J. Kreusch, A. Lallas, et al., Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks, JAMA Dermatol. 155 (1) (2019) 58–65.
[23] J.S. Birkenfeld, J.M. Tucker-Schwartz, L.R. Soenksen, J.A. Avils-Izquierdo, B. Marti-Fuster, Computer-aided classification of suspicious pigmented lesions using wide-field images, Comput. Methods Prog. Biomed. 195 (2020) 105631, doi:10.1016/j.cmpb.2020.105631.
[24] A. Xiang, F. Wang, Towards interpretable skin lesion classification with deep learning models, in: AMIA Annual Symposium Proceedings, vol. 2019, American Medical Informatics Association, 2019, p. 1246.
[25] K. Young, G. Booth, B. Simpson, R. Dutton, S. Shrapnel, Deep neural network or dermatologist? in: Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, Springer, 2019, pp. 48–55.
[26] C. Barata, M.E. Celebi, J.S. Marques, Explainable skin lesion diagnosis using taxonomies, Pattern Recognit. 110 (2021) 107413.
[27] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, S. Zhang, Ca-net: comprehensive attention convolutional neural networks for explainable medical image segmentation, IEEE Trans. Med. Imaging 40 (2) (2020) 699–711.
[28] D. Coppola, H.K. Lee, C. Guan, 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, 2020, pp. 734–735.
[29] A. Lucieri, M.N. Bajwa, S.A. Braun, M.I. Malik, A. Dengel, S. Ahmed, On interpretability of deep learning based skin lesion classifiers using concept activation vectors, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–10.
[30] A. Lucieri, M.N. Bajwa, A. Dengel, S. Ahmed, Explaining AI-based decision support systems using concept localization maps, in: International Conference on Neural Information Processing, Springer, 2020, pp. 185–193.
[31] J. Kawahara, G. Hamarneh, Fully convolutional neural networks to detect clinical dermoscopic features, IEEE J. Biomed. Health Inform. 23 (2) (2018) 578–585.
[32] D. Sonntag, F. Nunnari, H.-J. Profitlich, The skincare project, an interactive deep learning system for differential diagnosis of malignant skin lesions, arXiv preprint arXiv:2005.09448 (2020).
[33] A. Lucieri, A. Dengel, S. Ahmed, Deep learning based decision support for medicine–a case study on skin cancer diagnosis, arXiv preprint arXiv:2103.05112 (2021).
[34] E. Gibson, W. Li, C. Sudre, L. Fidon, D.I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, et al., Niftynet: a deep-learning platform for medical imaging, Comput. Methods Prog. Biomed. 158 (2018) 113–122.
[35] M.K. Hasan, S. Roy, C. Mondal, M.A. Alam, M.T.E. Elahi, A. Dutta, S.M.T.U. Raju, M.T. Jawad, M. Ahmad, Dermo-DOCTOR: a framework for concurrent skin lesion detection and recognition using a deep convolutional neural network with end-to-end dual encoders, Biomed. Signal Process. Control 68 (2021) 102661.
[36] S. Jiang, H. Li, Z. Jin, A visually interpretable deep learning framework for histopathological image-based skin cancer diagnosis, IEEE J. Biomed. Health Inform. 25 (5) (2021) 1483–1494.
[37] Data Language (UK) Ltd, Data language's explainable AI platform, 2021, (https://datalanguage.com/products/datalanguageai/explainable-ai-platform). Accessed: 2021-08-15.
[38] Decoded Health, The world's first clinical hyperautomation platform - a force multiplier for physicians, 2021, (https://www.decodedhealth.com/). Accessed: 2021-08-15.
[39] Hacarus Inc, Hacarus - sparse modeling based ai, edge ai with learning and inference capability, white box ai, 2021, (https://hacarus.com/). Accessed: 2021-08-15.
[40] A.B. Tosun, F. Pullara, M.J. Becich, D.L. Taylor, S.C. Chennubhotla, J.L. Fine, HistoMaprTM: an explainable AI (xAI) platform for computational pathology solutions, in: Artificial Intelligence and Machine Learning for Digital Pathology, Springer, 2020, pp. 204–227.
[41] T. Mendonça, P.M. Ferreira, J.S. Marques, A.R.S. Marcal, 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.
[42] J. Kawahara, S. Daneshvar, G. Argenziano, G. Hamarneh, Seven-point checklist and skin lesion classification using multitask multimodal neural nets, IEEE J. Biomed. Health Inform. 23 (2) (2019) 538–546, doi:10.1109/JBHI.2018.2824327.
[43] S.M.M. de Faria, J.N. Filipe, P.M.M. Pereira, L.M.N. Tavora, P.A.A. Assuncao, M.O. Santos, R. Fonseca-Pinto, F. Santiago, V. Dominguez, M. Henrique, Light field image dataset of skin lesions, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 3905–3908.
[44] R.V. Zicari, S. Ahmed, J. Amann, S.A. Braun, J. Brodersen, F. Bruneault, J. Brusseau, E. Campano, M. Coffee, A. Dengel, et al., Co-design of a trustworthy AI system in healthcare: deep learning based skin lesion classifier, Front. Hum. Dyn. 3 (2021) 40.
[45] 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, J. Am. Acad. Dermatol. 74 (6) (2016) 1093–1106.

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 Universitt Kaiserslautern (TUK), Germany and is also working as Research Assistant at Deutsches Forschungszentrum fõr Kõnstliche 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 concept-based explanation of skin lesion classifiers as well as the localisation of concept regions in input images.

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 Universitt 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 di-

agnosis 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 40 publications including several journal and high ranked conference papers.

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.

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.