International Journal of Data Science and Analytics
https://doi.org/10.1007/s41060-023-00401-z

REGULAR PAPER

Improving trust and confidence in medical skin lesion diagnosis through explainable deep learning

Carlo Metta[1] · Andrea Beretta[1] · Riccardo Guidotti[2] · Yuan Yin[3] · Patrick Gallinari[3] · Salvatore Rinzivillo[1] · Fosca Giannotti[4]

Received: 22 April 2022 / Accepted: 7 June 2023

Abstract
A key issue in critical contexts such as medical diagnosis is the interpretability of the deep learning models adopted in decision-making systems. Research in eXplainable Artificial Intelligence (XAI) is trying to solve this issue. However, often XAI approaches are only tested on generalist classifier and do not represent realistic problems such as those of medical diagnosis. In this paper, we aim at improving the trust and confidence of users towards automatic AI decision systems in the field of medical skin lesion diagnosis by customizing an existing XAI approach for explaining an AI model able to recognize different types of skin lesions. The explanation is generated through the use of synthetic exemplar and counter-exemplar images of skin lesions and our contribution offers the practitioner a way to highlight the crucial traits responsible for the classification decision. A validation survey with domain experts, beginners, and unskilled people shows that the use of explanations improves trust and confidence in the automatic decision system. Also, an analysis of the latent space adopted by the explainer unveils that some of the most frequent skin lesion classes are distinctly separated. This phenomenon may stem from the intrinsic characteristics of each class and may help resolve common misclassifications made by human experts.

Keywords Skin image analysis · Dermoscopic images · Explainable artificial intelligence · Adversarial autoencoders

1 Introduction

Artificial Intelligence (AI)-based decision support systems have recently gained a huge attention in different domains due to their remarkable performance. However, their adoption in sensitive scenarios that involves decision affecting humans, such as the medical one, has raised ethical concerns about the lack of transparency in decisions based on AI suggestions [1, 2]. There is a need to develop AI systems that can assist doctors in making informed decisions by supplementing their knowledge with information and suggestions from the AI system [3, 4]. However, if the reasoning behind the decisions of AI systems is not transparent, it would be difficult to achieve this goal. Skin image classification is a typical example of this problem. Indeed, when effective deep learning models are adopted to solve this problem, there are no indications that help in understanding the reasons for the decision outcome, which makes it difficult to have a natural interaction with the practitioner. Therefore, it is essential to augment currently adopted classification models with explainability components that enrich the interactions and provide the human with additional exploration and diag-

✉ Carlo Metta
carlo.metta@isti.cnr.it

Andrea Beretta
andrea.beretta@isti.cnr.it

Riccardo Guidotti
riccardo.guidotti@unipi.it

Yuan Yin
yuan.yin@sorbonne-universite.fr

Patrick Gallinari
patrick.gallinari@sorbonne-universite.fr

Salvatore Rinzivillo
salvatore.rinzivillo@isti.cnr.it

Fosca Giannotti
fosca.giannotti@sns.it

1 ISTI-CNR, Pisa, Italy

2 University of Pisa, Pisa, Italy

3 Criteo AI Lab, Sorbonne Université, Paris, France

4 Scuola Normale Superiore, Pisa, Italy

Published online: 21 June 2023

nosis tools [5]. This is the problem addressed in this paper when focusing on skin lesion diagnosis from images.

For all these reasons, eXplainable AI (XAI) has recently received much attention [2, 6, 7]. Saliency maps are the type of explanation most widely returned for image classifiers. A saliency map is an image highlighting each pixel's positive (or negative) contribution to the decision outcome. Various approaches are proposed in the literature to explain image classifiers through a saliency maps. As introduced in [6, 7], we underline that explanation methods can be categorized as *model specific or model agnostic*, depending on whether the explanation method exploits knowledge of the internal structure of the black box or not; *global or local*, depending on whether the explanation is provided for the black box as a whole or for any specific instance. Various model-specific explainers such as IntGrad [8], GradInput [9], and ε-LRP [10] are specifically designed to explain deep neural networks and return as explanation saliency maps. The saliency maps returned by these kinds of approaches are typically scattered and not easy to read in a critical medical situation. On the other hand, LIME [11] and SHAP [12] are two of the most well-known model and data agnostic local explainers. LIME randomly generates a local neighborhood "around" the instance to explain, labels them using the black box under analysis and returns an explanation using as surrogate model a linear regressor. SHAP leverages game theory and exploits the Shapley values of a conditional expectation function of the black box, providing for each feature the unique additive importance. LIME and SHAP can be applied to explain image classifiers and return explanation in the form of saliency maps. Unfortunately, both LIME and SHAP require a segmentation procedure that affects the explanation: the neighborhoods considered are no longer plausible instances but simply the image under analysis with some pixels "obscured" [13]. This is also not beneficial nor trustful in a medical context.

To overcome these issues, in [14] has been proposed ABELE,*Adversarial Black box Explainer generating Latent Exemplars*, a local model agnostic explainer specifically designed for image classifiers. Given an image to be explained and an image classifier, the explanation provided by ABELE is composed of (i) a set of *exemplar* and *counter-exemplar* images, and (ii) a *saliency map*. Exemplars and counter-exemplars are images classified with the same outcome as the input, and with a different outcome, respectively, while a saliency map highlights the areas which are more responsible for the decision.

The aim of this paper is to extend and exploit the methodologies illustrated in [3, 14, 15], and to study the usability of an explanation method into a real medical setting. In particular, we focus on skin lesion diagnosis from images. We rely on the labeled dataset available from the ISIC 2019 (*International Skin Imaging Collaboration*) image classification challenge. We train a state-of-the-art deep learning classifier using the ResNet [16] architecture on the dataset. After that, we explain the classifier decisions through ABELE [14]. In this way, the practitioner can easily reason on top of the exemplars and counter-exemplars returned by the explainer. Our goal is to assess to which extent these exemplar and counter-exemplar explanations are effectively useful through a user study involving humans. We design the experiment as a survey where participants are asked to address certain tasks on the basis of the classification outcome and the explanations. The novel contributions of this study w.r.t. existing ones consist of the following aspects. First, further refinement and evaluation of ABELE w.r.t. a real case study in the medical domain. Second, we introduce an analysis of the latent space of the adversarial autoencoder provided by ABELE. Third, we develop a user visualization module to analyze the explanations returned by ABELE. Fourth, we conduct a user study with domain experts, beginners, and unskilled people to assess the effectiveness of the explanations provided by ABELE. The results of the survey show that the usage of explanations increases trust and confidence in the automatic decision system. This phenomenon is more evident among domain experts and people with the highest level of education. Interestingly, the older segment of the population shows a chronic mistrust of AI models that is unaffected by the exposition of ABELE explanations. Also, we observe that after receiving wrong advice by an AI model, domain experts tend to decrease their trust in the same model for future analysis. As additional result, we highlight the analysis of the latent space of the autoencoder made available by ABELE. Furthermore, the latent space analysis suggests an interesting separation of the images that can hopefully be helpful in separating apart similar classes of skin lesion that are frequently misclassified by humans. Saliency maps produced by ABELE are found to be better than those produced by other local explainers, e.g. LIME and LORE [17].

The rest of the paper is organized as follows. Section 3 presents the methodology adopted. In Sect. 4 we illustrate the details of the case study addressed, and we present the visualization module. Section 5 illustrates the survey and shows the results obtained, while Sect. 7 presents the analysis of the latent space. Finally, Sect. 8 summarizes the contribution and proposes future research directions.

2 Related work

XAI has gained significant attention in medical imaging. It aims to provide a transparent and interpretable understanding of the underlying processes and decisions made by AI models. In medical imaging, XAI can play a crucial role in improving the trust and confidence of healthcare providers and patients in AI-based diagnoses and treatments. Several

studies have evaluated XAI in medical imaging, including chest X-rays [18], CT scans [19], and MRI scans [20]. These studies have applied various XAI methods, such as saliency maps, attribution maps, and decision trees. One of the earliest studies in this area is the work by Jampani et al. [21], which applied saliency map models on different medical image domains. Since then, many studies have also explored the use of decision trees and rule-based systems to provide explanations for AI-based diagnoses in medical imaging. For example, Seung et al. [22] used decision trees to explain the predictions made by a deep learning model on chest X-rays. In [22] is shown that decision trees can effectively provide an interpretable explanation of the AI model's decision-making process.

Furthermore, there have been several evaluations of XAI in the context of medical imaging diagnosis, including breast cancer diagnosis [23], lung nodule diagnosis [24], and brain tumor diagnosis [25]. Overall, these studies highlight the importance and novelty of XAI in medical imaging, as they demonstrate the potential for XAI methods to increase trust and confidence in AI-based diagnoses and to provide a better understanding of the underlying processes used by AI models.

XAI in medical imaging still has several issues, problems, and challenges that need to be addressed [26], such as lack of trust, data bias, interpretability, privacy concerns and integration with clinical workflow.

This work aims to provide a human-centered approach to address ongoing challenges in XAI in medical imaging and contribute to the adoption of these algorithms in clinical practice by building reliable and transparent decision support tools to integrate XAI methods into the clinical workflow.

Our XAI methodology falls under the category of Generative Explanation-based Methods, where a generative process is exploited to create visual explanations. The Contrastive Explanations Method (CEM) [27] generates explanations that display the minimum regions in a given image that must be present or absent to justify a particular classification decision. Other works [28–30] generate explanations that emphasize the features that should be altered to increase or decrease the classifier's confidence in the prediction (i.e., prototypes or counterfactuals). Explanation by Progressive Exaggeration [31] proposes a method for explaining the outcome of a black box classifier by leveraging over a Generative Adversarial Network (GAN) [32] and gradually changing the input query in a way that changes the model's prediction: the method is model agnostic and only requires access to the predictor values and its gradient with respect to the input. Our contribution aligns with recent advancements in the field. However, we choose to utilize a different architecture, the Adversarial Autoencoder (AAE). One advantage of AAE over GAN is that AAE has more precise control over the latent space, which is the representation of the data in a lower-dimensional form. This allows AAE to generate samples that are closer to the real data distribution and to ensure that the generated data is coherent and meaningful. Another advantage of AAE over GAN is that AAE can be trained with a reconstruction loss, which ensures that the generated data is similar to the input data. In contrast, GANs are trained based on a min-max game between the generator and the discriminator, which makes it difficult to ensure that the generated data is similar to the real data. Additionally, AAEs can be used for unsupervised representation learning, where they can learn a compact representation of the data that can be used for other tasks, such as classification. This is because the encoder part of the AAE learns to map the data to the latent space, and the decoder part learns to map back from the latent space to the original space.

In conclusion, AAE provides a more controllable and interpretable method for generating new data compared to GANs, making it a more useful and robust tool for XAI.

3 Methodology

In this section, we briefly present the two main components of the methodology adopted to classify and explain the dataset. Details can be found in [3, 14, 16].

3.1 ResNet classifier

In order to provide a black box classifier that performs sufficiently well for downstream learning steps, we choose to train a neural network with an architecture powerful enough to accomplish image classification. In particular, we selected a *ResNet*, a popular architecture providing validated performance on many complex datasets and tasks [16]. Instead of training the ResNet from scratch, we choose to perform a *transfer learning* task with a ResNet pre-trained on the ImageNet dataset. This training strategy is largely applied for the case where the number of data is limited with respect to the complexity of a network [33]. To perform the transfer learning, we replace the last fully connected layer with the newly initialized one. The number of output dimensions is adapted to the number of classes in the dataset. Then, the classification layer is learned from scratch, and the rest of the ResNet is fine-tuned. As loss function, we adopt a binary cross entropy loss for each class, so that the task can be considered as individual one-vs-rest binary classification problem.

3.2 ABELE explainer

ABELE is a local model agnostic explainer that takes as input an image x and a black box classifier b, and returns *(i)* a set of *exemplar* and *counter-exemplar* images, and *(ii)* a *saliency map*. Exemplars and counter-exemplars are images synthet-

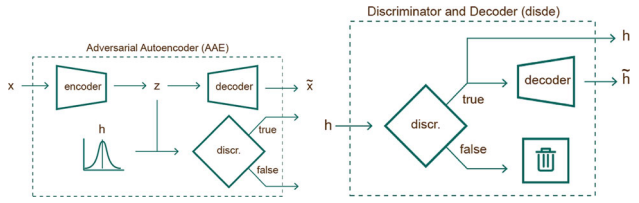

Fig. 1 AAE Architecture, Discriminator and Decoder modules, [14]

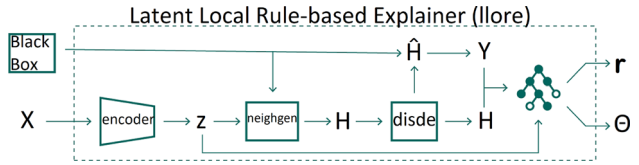

Fig. 2 Latent local rules extractor (LLORE) module, [14]

ically generated and classified with the same outcome as x, and with an outcome other than x, respectively. They can be visually analyzed to understand the reasons for the decision. The saliency map highlights the areas of x that contribute to its classification and areas that push it into another class.

In short, ABELE works as follows. First, it generates a neighborhood H in the latent feature space exploiting an Adversarial Autoencoder (AAE) [34]. The AAE architecture (Fig. 1) includes an `encoder`: $\mathbb{R}^n \rightarrow \mathbb{R}^k$, a `decoder`: $\mathbb{R}^k \rightarrow \mathbb{R}^n$ and a `discriminator`: $\mathbb{R}^k \rightarrow [0, 1]$ where n is the number of pixels in an image and k is the number of latent features. The image $x \in \mathbb{R}^n$ to be explained is passed as input to the autoencoder where the `encoder` returns the latent representation $z \in \mathbb{R}^k$ using k latent features with $k \ll n$. The neighborhood generation of H (`neighgen` module in Fig. 2) may be accomplished using different strategies. In our experiments, we adopt a genetic approach maximizing a fitness function like in [35]. The module in Fig. 2 is named LLORE, *Latent Local Rules Extractor*, as a latent variant of LORE [35]. After the generation process, for any instance $h \in H$, ABELE exploits the `disde` module (Fig. 1) for both checking the validity of h by querying the `discriminator` and decoding it into $\widetilde{h}$. Then, it queries the black box b with $\widetilde{h}$ to get the class y, i.e. $b(\widetilde{h}) = y$. Given the local neighborhood H, ABELE builds a decision tree classifier c trained on H labeled with $b(\widetilde{H})$. The surrogate tree aims to imitate the behavior of b in the neighborhood H. It extracts the decision rule r and counterfactual rules Φ to create *exemplars* and *counter-exemplars*. Figure 2 depicts the process of creating the decision tree, starting from the image to be explained and resulting in the extraction of decision and counterfactual rules.

The overall effectiveness of ABELE lies in the goodness of the encoder and decoder function adopted: the better is the autoencoder, the more realistic and useful will be the explanations. In the next section, we highlight some peculiarities of the structure of the autoencoder required to obtain reliable results for the ISIC dataset.

3.3 Progressive growing adversarial autoencoder

We summarize here the customization of ABELE we carried on in order to make it usable for complex image classification task. Details can be found in [3]. Generative Adversarial models are generally not easy to train as they are usually affected by a number of common failures. These problems vary from a diversified spectrum of *failures in convergence* to the famous *mode collapse* [36], the tendency by the generator network to produce a small variety of output types. Such problems mainly arise from the competing scheme generator and discriminator are trained on. In addition, we often face the further complication to deal with real world datasets that are far from ideal: fragmentation, imbalance, lack of uniform digitization, shortage of data are primary challenges of big data analytic for healthcare. Training an AAE in a standard fashion to reproduce samples from ISIC dataset without taking special care of all issues mentioned above resulted in extremely poor performance, mostly due to a persistent collapse mode.

In order to overcome such generative failure and dataset limitations, we implemented a collection of cutting edge techniques that altogether are capable of addressing all the issues we mentioned and successfully training an AAE with adequate performance. In particular, we address mode collapse using ad hoc tricks like Mini Batch Discrimination [37] and Denoising autoencoders [38]. As model, we implemented a Progressive Growing Adversarial Autoencoder. Progressive Growing GANs [39] have been introduced as an extension of GANs. Progressive growing helps to achieve a more stable training of generative models for high resolution images like in our case. The main idea is to start with a very low resolution image and step by step adding block of layers that simultaneously increase the output size of the generator model and the input size of the discriminator model until the desired size is achieved. However, while in a GAN, the discriminator is linked to the generator output, in an AAE, the discriminator takes as input the encoded latent space instead of the full reconstructed image. Thus, in [3] we define Progressive Growing Adversarial Autoencoder (PGAAE) as follows. Starting with a single block of convolutional layers for encoder and decoder, we are able to reconstruct low resolution images (14×14 pixels), then step by step we increase the number of blocks until the networks are powerful enough to manage images of the desired size, i.e., 224×224 pixels in our case. The latent space dimension is kept fixed, consequently the discriminator takes as input tensors always of the same size. Although one could fix also the network of the discriminator, we found helpful to progressively increasing also the width of this network so that

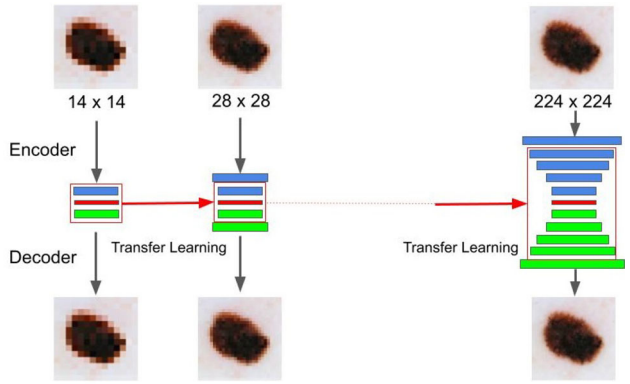

Fig. 3 A progressive growing AAE. At each step, an autoencoder is trained to generate an image that is twice the size of the previous one, starting from an image of 14x14 pixels and gradually increasing to an image of 224x224 pixels. The learned features from one autoencoder are then transferred to the next. To handle the growing image size, both the encoder and decoder networks are expanded by adding one convolutional block at each step. The transfer learning is confined to the shared network architecture

the discriminator can deal each step with a more structured information. The incremental addition of the layers allows the autoencoder to first learn large scale structure and progressively shift the attention to finer detail.

The network scheme is reported in Fig. 3. In our implementation, we used five blocks of layers in order to have a trained network able to reproduce skin lesion images of size 224×224 pixels. In summary, as learning tricks we relayed on minibatch discrimination, denoising and progressive structure. Mode collapse is greatly reduced, and we are able to generate variegate and good quality synthetic skin lesion images with ABELE acting as exemplars and counter-exemplars.

We highlight the fact the ABELE pipeline works smoothly regardless of the chosen classifier. Of course, different classifiers would produce different explanations.

4 Case study: skin lesion diagnosis

The case study and the properties of the training dataset are described in this section, along with the details of training the black box classifier and the autoencoder.

4.1 Dataset

The International Skin Imaging Collaboration (ISIC), sponsored by the International Society for Digital Imaging of the Skin (ISDIS) proposed the *skin lesion analysis towards melanoma detection challenge* to improve the international effort in melanoma diagnosis. [1] The challenge consists in developing a classifier to recognize among nine different diagnostic categories of skin cancer: MEL (Melanoma), NV (Melanocytic nevus), BCC (Basal cell carcinoma), AK (Actinic keratosis), BKL (Benign keratosis), DF (Dermatofibroma), VASC (Vascular lesion), SCC (Squamous cell carcinoma), UNK (Unknown, none of the others/out-of-distribution). The dataset is composed of a training set of 25,331 images of skin lesions and their category (labels); a test set of 8238 images of which the label is not publicly accessible.

4.2 ResNet training

We separated the training into two parts: 80% samples used for training and 20% for validation. The UNK category was not taken into account for training purposes. We made this decision based on the task we considered, i.e., to train a classifier for diagnostic purposes. Indeed, the classifier should make reliable decisions to help human doctors. A reliable model should reject UNK samples. Thus, we give our model an unsupervised rejection choice by adding an extra output corresponding to a UNK class even during training, to let the model reject a sample when it is not confident enough. In particular, a sample is rejected if the neural network gives UNK the highest output over all other classes. This consideration is indispensable from a medical diagnostics viewpoint. It would be better to reject an out-of-distribution sample and leave the decision to human doctors than giving it some arbitrary wrong label. Also, since images have different resolutions, we applied the following preprocessing:

- For the training, the images are randomly rescaled, rotated and cropped to generate the input to the network.[2] Resolution of the preprocessed images is 224×224.
- For the validation and test, each image is firstly rescaled to 256×256 according to the shorter edge, then cropped at the center into a 224×224 image.

For the evaluation, we employed the same metric adopted in the submission system of the original challenge. The model is evaluated on (our) test set with the normalized (or balanced) multi-class accuracy defined as the average of recall obtained in each class. This metric makes all the classes equally important to avoid that classifier performs well only for dominant classes. The trained ResNet model, which is the best model evaluated on the validation set mentioned above, achieves 0.838 of balanced multi-class accuracy on the test set. Also, since the images are captured under controlled conditions, there is no significant distributional shift in the shuffled data samples used to create the training set and the validation sets, so the difference between them is small

[1] https://challenge2019.isic-archive.com/.

[2] Such preprocessing does not deform the lesions in the image.

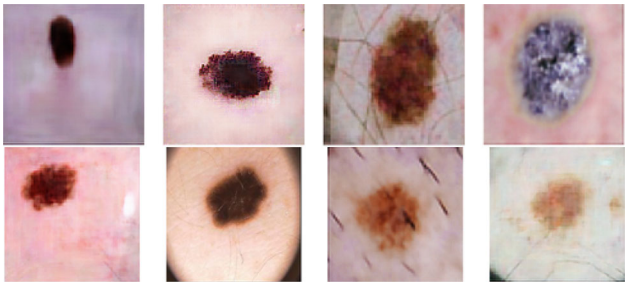

Fig. 4 Synthetic skin lesion samples generated by ABELE and classified as Melanocytic Nevus by the ResNet black box, except for the upper right image classified as Actinic Keratosis

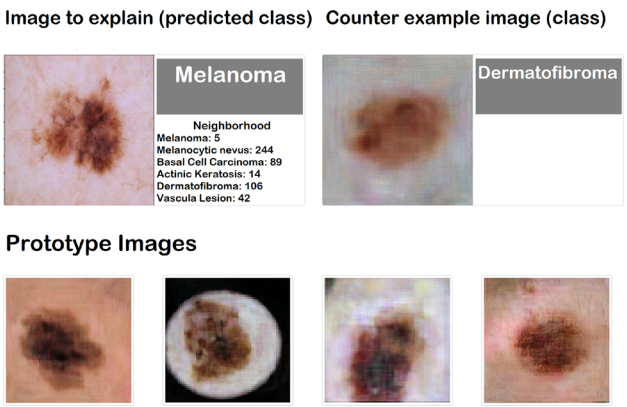

Fig. 5 User visualization module to present the classification and the corresponding explanation. The upper part presents the input instance and a counter-exemplar. The lower part shows for exemplars that share the same class as the input

and the cross-validation is not necessary. Since the amount of data is small, we fine-tune an off-the-shelf pre-trained ResNet on our dataset to avoid overfitting. The architecture is chosen beforehand according to its past performance on the image classification task and its computational complexity, not according to its performance on our dataset. The learning rate is selected from a sparse grid and set to 10^{-4} according to the convergence speed and accuracy on the validation set. Finally, we keep the model with the best score on the validation set.

4.3 PGAAE training

A customization of ABELE was necessary in order to make it usable for the complex image classification task addressed by the ResNet black box classifier. Details are available in [3]. After a thorough fine tuning of all three networks structures (encoder, decoder and discriminator), our PGAAE with 256 latent features achieves a reconstruction error measure through *root mean square error* that ranges from 0.08 to 0.24 depending on whether we consider the most common or the most rare skin lesion class. We selected 256 as number of latent features because, from preliminary experiments, it was the number that simultaneously guaranteed a good reconstruction error, a good resolution of the images and did not involve an excessive waste of computational resources. Also, for images of the desired size, i.e., 224×224, it is common in the literature to choose a number of latent features that varies between 64 and 512. Data augmentation was necessary to overcome scarcity and imbalance of the dataset. Mode collapse was greatly reduced, and we were able to generate variegate and good quality skin lesion images (Fig. 4). ABELE equipped with PGAAE, can produce meaningful explanations, as demonstrated in a survey of real participants, which is discussed in the following section.

4.4 ABELE user visualization

Module

We present here the novel visualization module for the explanations returned by ABELE. The module helps users understand both the suggestions made by the black box model and the explanations produced by ABELE. Figure 5 shows a screenshot from a web application[3] we implemented to present to the user the outcome of our system on a specific instance. The application is divided into two sections, with the upper section showing the analyzed image, the classification made by the black box, and a synthetic counter-exemplar image generated by ABELE. In Fig. 5, we have an example of an image x of *Melanoma* and a similar synthetic image classified as *Dermatofibroma* as a counter-exemplar. The neighborhood generated by ABELE, represented as a list, provides a glimpse into the diverse instances in the latent space surrounding the analyzed image. The counter-exemplar, i.e. an image classified differently from the original instance, is selected among this list as the image that minimizes the euclidean distance with x (over the latent space) and maximizes the classification prediction but w.r.t. a different label. Finally, the bottom section of the module displays four exemplars, a set of images generated by ABELE that have the same label assigned by the black box to x.

The ABELE visualization module is implemented in Javascript as a web application. It communicates with a backend that exposes the functionalities of the black box and of ABELE by means of a RESTful interface. We implemented and deployed a demonstrator of the system by letting the user choose from a set of instances, instead of uploading a new one. This demonstrator has been used for the preparation of the survey presented in the next section.

[3] https://kdd.isti.cnr.it/isic_viz/

5 Validation and survey

We designed a survey to assess the impact of ABELE explanations on skin lesion diagnosis. The primary goal is to determine the usefulness of explanations in supporting doctors and medical experts in diagnosing and treating skin cancers, and to evaluate their confidence in black-box-based diagnosis models and the explanations provided by the explainer.

5.1 Survey structure

The survey is composed of ten questions, each presenting a different medical image case. Each question is organized into four different points, following the same structure for the ten cases. We denote each question with its progressive number, i.e., Qi with $i \in [1, 10]$.

Point 1 (P1). The participants are presented with an unlabeled skin lesion image randomly chosen among the dataset and its explanation as generated by ABELE and presented by the visualization module. In particular, we presented two exemplars and two counter-exemplars of another lesion class to the user. Participants were asked to classify the given image among two different given classes exploiting the explanation. This point aims to understand if the explanations returned by ABELE significantly help in separating different images, even for non-expert users. From another perspective, this can be considered the human evaluation of the *usefulness* metric synthetically observed in [14].

Point 2 (P2). The participants are presented with a labeled image, and they are asked to quantify their level of confidence in the black box classification (using a 0-100 slider).

Point 3 (P3). The participants are presented with the same labeled image of P2, but this time with the visual aid of the explanation returned by ABELE, and they are asked to quantify their confidence once more after looking at the explanations. The objective of points P2 and P3 is to understand if there is an increase/decrease in the confidence towards the AI after having observed an explanation.

Point 4 (P4). The participants are asked to quantify how much the exemplars and counter-exemplars helped them to classify skin lesion images in accordance with the AI, and how much they trust the explanations produced by ABELE.

During the survey, participants were not informed of the correctness of their prediction, nor they received further suggestions by looking back at their previous answers or explanations. To investigate the recipient's reaction to incorrect advice, in Q6 we intentionally entered a wrong classification (P2) followed by further wrong advice concerning exemplars and counter-exemplars (P3). All the other nine instances were selected among correctly classified cases.

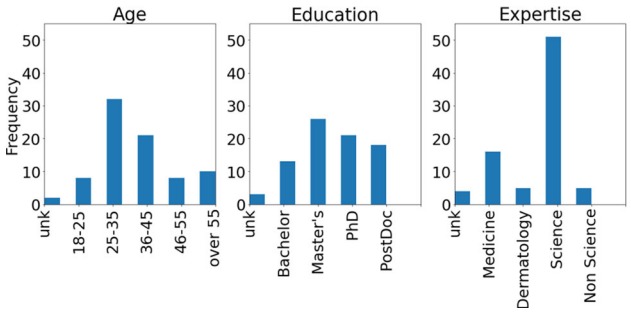

Fig. 6 Demographic statistics of the survey participants

5.2 Hypothesis and objectives

The structure of the investigation reflects the following hypothesis we intend to address.

- *H1:* The explanations returned by ABELE help the recipients in the classifications task, especially domain experts who are supposed to achieve a higher classification score (implicit assessment through P1).
- *H2:* The explanations returned by ABELE improve the recipients' trust and confidence toward the black box classification (implicit assessment through P2, and P3, explicit through P4).
- *H3:* After receiving the wrong advice from the black box, participants show a substantial decline in confidence and trust toward that model (implicit assessment through the error inserted).

5.3 Survey results

A total of 156 participants completed the study. Participants signed up for the survey online after digitally signing a consent form, followed by a short demographic survey and a brief introduction about all the different types of skin lesion cancer involved in the process. Since participants were not forced to answer all ten questions, we chose to consider only those who had completed at least 10% of the questionnaire, i.e., at least an entire question with an answer for each of the four points. Aggregate demographic statistics of the participants are available in Fig. 6. Of the participants, 94% have a scientific background, with 27% of them having completed studies in medicine or dermatology.

First, we have attached to each participant a score that measures their performance in P1 to test their ability to classify skin lesion images by exploiting the explanation. We divided participants into two sub-samples. *Sub-sample A* contains participants who achieved a score of at least 70% images correctly classified, and *sub-sample B* contains the remaining participants. Although the average performance

Table 1 Participants' confidence in the classification of the black box before and after receiving the explanation of ABELE

	Expert		Non expert	
	Before	After	Before	After
Q1	75.6	**87.0**	69.3	**84.3**
Q2	65.8	**73.0**	56.9	**60.8**
Q3	70.8	**80.0**	**63.2**	59.9
Q4	72.6	**83.8**	60.5	**80.5**
Q5	95.2	**98.4**	89.2	**91.2**
Q6	39.2	**61.8**	54.4	**76.3**
Q7	94.0	**96.2**	81.1	**89.7**
Q8	55.8	**62.8**	76.3	**84.0**
Q9	**59.2**	57.2	60.3	**62.2**
Q10	69.4	**87.6**	63.5	**81.0**

The values with the highest confidence are highlighted in bold

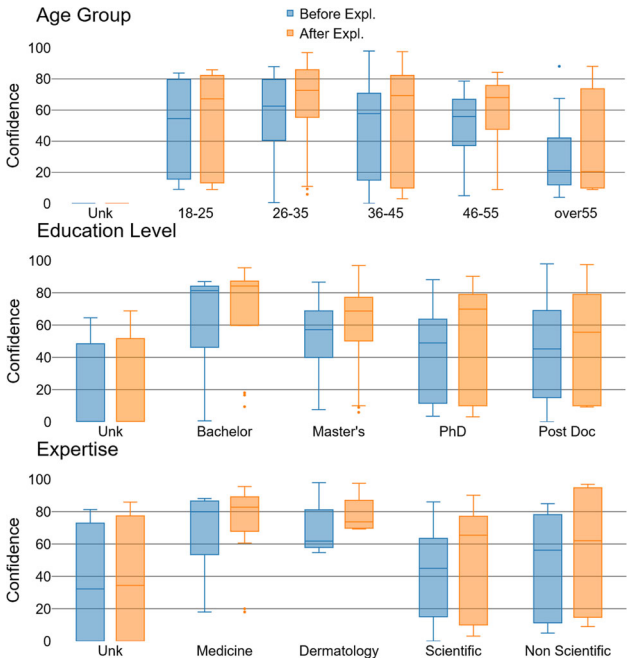

Fig. 7 Participants' confidence among different age groups (top), education level (center), domains (bottom), before and after explanations

was remarkable (score 82.02%) among all participants, including those without domain expertise or specific medical knowledge (score 78.67%), we are interested in the sub-sample of people specialized in medicine or in dermatology (score 91.26%). The results of the one-way ANOVA on ranks (Kruskal-Wallis H test) [40] applied to sub-samples A and B showed no significant impact on classification performance based on either education level or age distribution. However, it shows a pronounced effect ($F = 4.061$, $p = 0.043$) given by the specific field in which the participants specialize—participants with a medicine degree and/or a dermatology specialization are more frequent among people in sub-sample A. Thus, H1 holds for domain experts, and even participants from other domains demonstrated noteworthy performance.

Table 1 shows the participants' confidence in the black box classification before and after looking at the explanations returned by ABELE, i.e., reports the responses obtained for P2 and P3. Except for Q3 and Q9, all the other questions show a significant increase of trust after looking at exemplars and counter-exemplars, i.e., an increase between P2 and P3, indicating that, in general, the explanations help in increasing the model trust. However, Table 1 suggests that Q3 anomaly was caused by the sub-sample of non-medical expert. Such an increase in confidence from 67.69% to 77.12% is maximum for Q6 (+21.95%), the only one which was misclassified by the black box. Moreover, Q6 presents a confidence, prior to the explanations, lower than all the others, i.e., of 53.08%. The reason for this phenomenon could be a significant decrease in participant's confidence after receiving incorrect advice, which is fully restored if they receive additional consistent incorrect suggestions. Participants reject incorrect advice, but they tend to adapt and reset their understanding if they consistently receive erroneous suggestions.

Increase in confidence was not uniform among all participants. The results summarized in Fig. 7 seem to suggest the following aspects. First, there is a not negligible increase in confidence among all ages except for age group over 55, for which not only the confidence is very low in itself but even decreases after having benefited from the explanations. This may be caused by the fact that the older segment of the population has an inherent distrust of AI models in general, while younger sections of the population are mentally more open to such models. Second, the confidence before looking at the explanations decreases as the level of study increases, while more educated participants show a notable increase after the explanations (a possible reminiscence of the Dunning–Kruger effect [41]). Third, as expected, the confidence level is much higher for people belonging to the medical domain than for participants from other scientific disciplines and even more so for those specializing in non-scientific disciplines.

As mentioned earlier, Q6 was specifically chosen from those misclassified by the black box, in order to investigate participants' reaction and behavior in that and subsequent instances (H3). The results show a slight mistrust toward the sixth black box classification, although there is no statistically significant drop in confidence after receiving wrong advice by an AI model (68.75% for Q1 to Q5, 60.03% for Q6 and 66.71% for Q7 to Q10). On the contrary, if we restrict our study to the sub-sample of medical experts, Fig. 8 shows a 14% drop of confidence after receiving the wrong advice (78.04% for Q1 to Q5, 56.19% for Q6 and 63.95% for Q7 to Q10), supporting H3: after receiving wrong advice from an

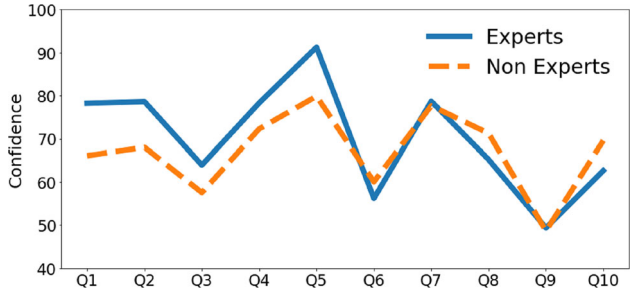

Fig. 8 Participants confidence towards ABELE explanations

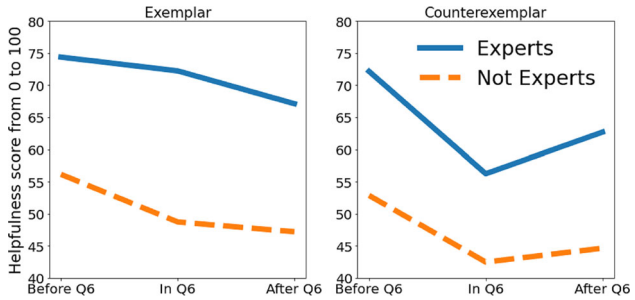

Fig. 9 How much exemplars and counter-exemplars helped according to the participants' responses, divided between groups of experts and non-experts

AI model, domain experts show a decline in confidence and trust toward that model in subsequent instances.

Figure 9 summarizes how exemplars and counter-exemplars impacted participant recognition of lesion classes, as reported by respondents in P4. The trend observed in analyzing confidence before and after Q6 is consistent, with both experts and non-experts showing a notable decline in confidence in the help provided by exemplars and counter-exemplars. As expected, ABELE's explanations are more helpful for medical experts than for the general population. Additionally, exemplars were more effective than counter-exemplars for both experts and the general population. This behavior may be due to the 8-class classification task, as in a binary classification task, the significance of exemplars and counter-exemplars may be similar, but as the number of classes increases, the importance of exemplar tends to increase.

6 Comparison of saliency maps

In Fig. 10, we report the explanation of ABELE for an example. Synthetic exemplar and counter-exemplar are much more informative than standard saliency maps. Saliency maps can be compared to those produced by existing explainers, e.g., LIME and LORE. The saliency maps shown in Fig. 11 have a deletion AUC (Area Under Curve) score ([42]), respectively, of 0.888 (LIME), 0.785 (LORE) and 0.593 (ABELE). The deletion metric measures the drop in the probability of a specific class as important pixels (as rated by the saliency map) are progressively removed from the image. A small area under the curve is indicative of a good explanation. We performed the calculation of the deletion metric for a set of 200 sample images and then calculated the average of the scores. As expected, segmentation based methods struggle in providing meaningful saliency maps (LIME: 0.736 mean AUC score, LORE: 0.711 mean AUC score), while ABELE generates more granular maps (0.461 mean AUC score). In Fig. 12 (Top), we report the deletion curves expressed as mean AUC of accuracy vs percentage of removed pixels for 200 sample images. We observe that ABELE deletion curve drops more rapidly and at an earlier stage relative to the percentage of removed pixels, indicating that the saliency map is finer and more granular.

A similar pattern arises when considering an insertion metric ([42]), which takes a complementary approach. It evaluates the effect of the black box prediction by incrementally adding each pixel in order of increasing importance. Thus, we expect that the black box performance increases by adding more and more features, resulting in monotonically

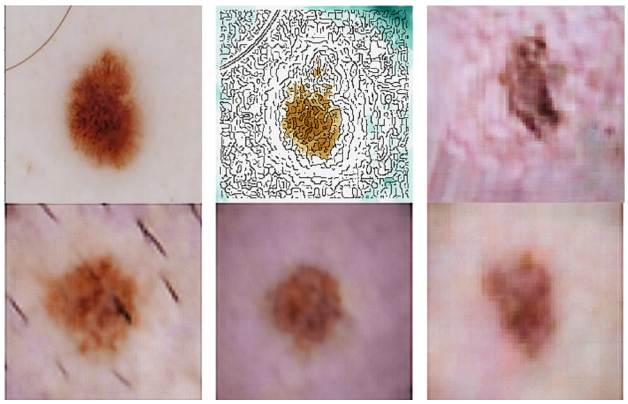

Fig. 10 ABELE explanation for a Melanocytic Nevus consisting in: original image (top left), saliency map respecting the latent rule (top center), counter-exemplars of a Basal Cell Carcinoma (top right) and three different exemplars (bottom)

Fig. 11 Saliency maps for LIME (left), LORE (center) and ABELE (right). LIME and LORE highlight the macro-regions of the image that contribute positively (green) or negatively (red) to the prediction, while ABELE provides a more fine-grained level of information with a divergent color scale, from relevant areas (dark orange) to low-significant areas (green/cyan)

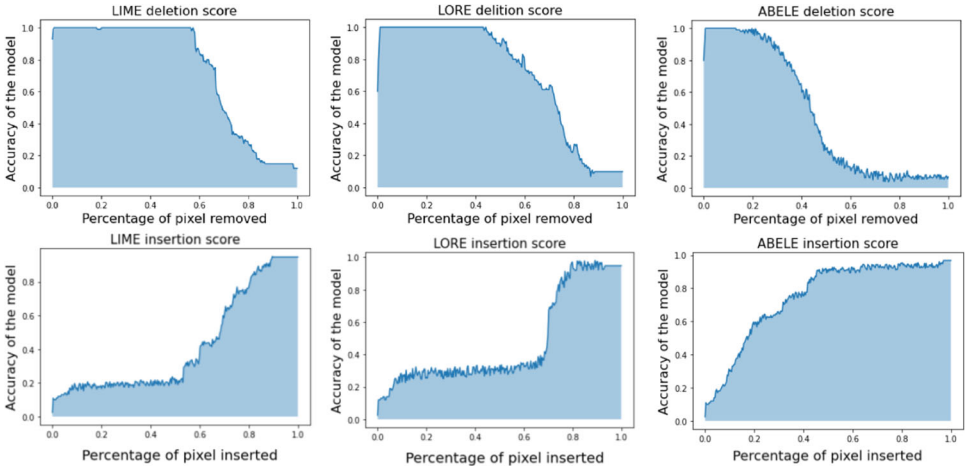

Fig. 12 Deletion (Top) and Insertion (Bottom) metrics expressed as mean AUC of accuracy vs percentage of removed or inserted pixels for 200 sample images. ABELE deletion curve drops earlier and faster relative to the percentage of removed pixels, signaling finer and more granular maps. ABELE insertion curve grows much earlier respect to LIME and LORE

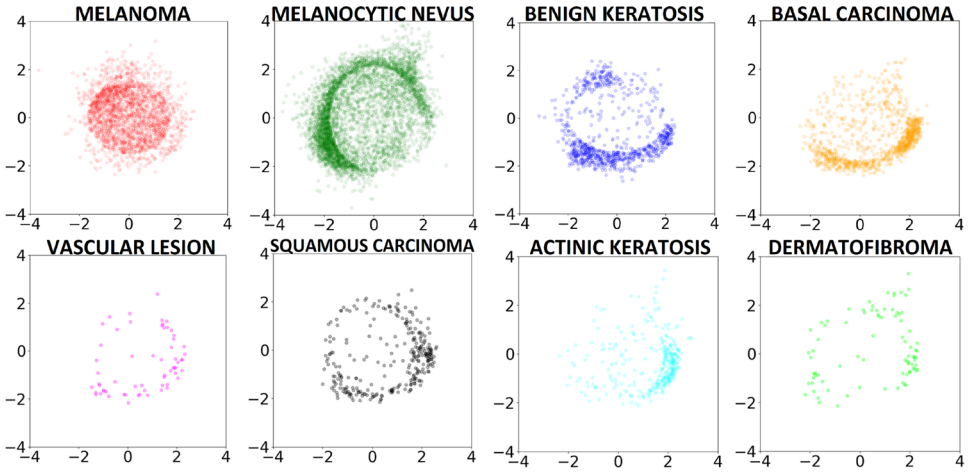

Fig. 13 Training set represented in two dimensions through a MDS applied on the latent space learned by the PGAAE

increasing model performance. A bigger area under the curve refers to a better explanation. Figure 12 (Bottom) reveals that ABELE shows a consistently better insertion score for an average of 200 samples (0.417 (LIME), 0.471 (LORE) and 0.748 (ABELE)). ABELE insertion curve grows rapidly, indicating that the saliency map captures more accurately the portions of the image that are most important to the classifier.

7 Explaining through latent space analysis

The PGAAE trained for ABELE projects the ISIC dataset into a 256-dimensional latent space with a coherent posterior distribution. Indeed, as shown in [14], as a side effect of ABELE, we can exploit the latent space to visualize the level of proximity of individual instance of the dataset and gain useful insights. In particular, we believe that such visual aid can help medical expert and data scientist to better understand different skin cancer characteristics and exploit it to further improve the classification performance, or the trust in the explainer.

We adopt a Multidimensional Scaling (MDS) [43] as a form of dimensionality reduction to translate information about pairwise distances among latent projections into a configuration of the same cardinality mapped into a Cartesian plane. Thus, through MDS we turn the latent space with 256 dimension into a visual space with 2 dimensions. Figure 13 shows the latent encoding of 8 skin cancer classes. We observe that some primary features of skin lesion can be also retrieved from such 2D projection. Indeed, from Fig. 13 it is clear that all skin lesion classes except *Melanoma* tend to avoid the center of each diagram and accumulate over a circle, while *Melanomas*, the most dangerous skin cancer, reside in the center of the plot.

It can be argued that such behavior is related to the similarity between these skin cancer classes. In [44], authors state that *Benign Keratosis* is one of the lesions for which melanoma is commonly misdiagnosed; this error occurred in 7.7% to 31.0% of cases, depending on the study. To delve deeper into the differentiation capabilities between different skin lesion classes, we decided to train a Random Forest (RF) classifier [45] with 500 estimator trees over the 2D MDS space. The RF classifier is able to separate apart *Melanoma* from *Benign Keratosis* with 85.60% accuracy (see Fig. 14-*left*). The Random Forest classifier assists

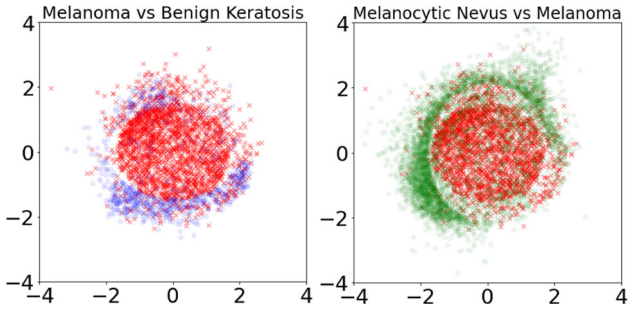

Fig. 14 Visual separation between Melanoma and Benign Keratosis (Left) and Melanocytic Nevus (Right)

the user in visually separating different skin lesions and is less complex than the original black box while maintaining comparable performance. Another important class is indeed *Melanocytic Nevus* that shows peculiar features that can be partially justified by the odd representation in Fig. 13. Here many samples still reside at the center of the plot: from 30% to 50% of all melanomas and more than half of those in young patients evolve from initially benign nevi [46]. The RF classifier trained over the 2D-space is also able to separate *Melanoma* from *Melanocytic Nevus* with 78.53% accuracy (see Fig. 14-*right*). These performances are comparable with the original black box accuracy for the overall scores in the ISIC 2019 challenge [3], and state-of-the-art classification accuracy with deep convolutional neural network [47].

Nowadays, the detection of melanomas is one of the most researched topics in the oncologic domain [48–50]. Predicting the transformation of a nevus into a malignant melanoma remains a challenging task for both clinicians and computers. To address this issue, future research needs to consider the evolution of oncologic data over time. Our methods and findings may assist clinicians in accurately evaluating the potential for a benign skin lesion to become melanoma.

8 Conclusion

In this paper, we have shown how it is possible to instantiate methodologies of classification and post hoc explanation in a real case study for skin lesion detection. In particular, we have proved that, after being customized and trained carefully, ABELE is able to produce meaningful explanations that can significantly help practitioners. Such visual explanations are more informative and qualitative better than those produced by other existing local explainers. The non-trivial and time-consuming step is the training of the generative model. The latent space analysis suggests an interesting repartition of image over the latent space, and it can hopefully be helpful in separating apart similar classes of skin lesions that are frequently misclassified by humans (benign from malignant). Also, we conducted a survey involving real user experts and not experts in skin cancer and of healthcare domains. The survey supports the hypothesis that explanation methods without consistent validation are not useful. As future research directions, it will be interesting to apply ABELE explainer to different diseases and health domains, especially domains where the only meaningful data is the raw image or scan of a particular internal body portion. Indeed, in skin lesion cancer also the touch plays an important role in doctors' analysis, not just the image. Also, we would like to extend the user visualization module presented to a real-time explanations generator. This improvement would require significant efforts and resources as in the current implementation extracting explanations can be time-consuming depending on the image.

Acknowledgements This work is supported by the European Community under the Horizon 2020 programme: G.A. 871042 *SoBigData++*, G.A. 952026 *HumanE AI Net*, G.A. 101092749 *CREXDATA*, ERC-2018-ADG G.A. 834756 *XAI*, G.A. 952215 *TAILOR*, and the NextGenerationEU programme under the funding schemes PNRR-PE-AI scheme (M4C2, investment 1.3, line on AI) FAIR (Future Artificial Intelligence Research), and "SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Prot. IR0000013.

Author Contributions The contribution of each author was as follows: CM, RG, SR: Conceptualization, Methodology, Software, Visualization, Validation; AB: User Study, Ethical Consent Preparation; YY: Software, Validation; FG, PG: Conceptualization, Funding acquisition.

Funding Open access funding provided by ISTI - PISA within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors have no competing interests relevant to the content of this article.

References

1. Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., Turini, F.: Meaningful explanations of black box AI decision systems, in *AAAI*. AAAI Press, pp. 9780–9784 (2019)
2. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)

3. Metta, C., Guidotti, R., Yin, Y., Gallinari, P., Rinzivillo, S.: Exemplars and counterexemplars explanations for image classifiers, targeting skin lesion labeling, in *IEEE ISCC*, (2021)
4. Panigutti, C., Perotti, A., Pedreschi, D.: Doctor XAI: an ontology-based approach to black-box sequential data classification explanations, in *FAT**. ACM, pp. 629–639 (2020)
5. Markus, A. F., Kors, J. A., Rijnbeek, P. R.: The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies, *J. Biomed. Inform.*, 113, (2021)
6. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52 138-52 160 (2018)
7. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 931–9342 (2019)
8. Sundararajan, M., et al.: Axiomatic attribution for dnn, in *ICML*. JMLR, (2017)
9. Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not just a black box: learning important features through propagating activation differences, *CoRR*, vol. https://arxiv.org/abs/1605.01713, (2016)
10. Bach, S., Binder, A., et al.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**(7), e0130140 (2015)
11. Ribeiro, M. T., Singh, S., Guestrin, C.: Why should I trust you?: explaining the predictions of any classifier, in *KDD*. ACM, pp. 1135–1144 (2016)
12. Lundberg, S. M., Lee, S.: A unified approach to interpreting model predictions, in *NIPS*, 4765–4774 (2017)
13. Guidotti, R., Monreale, A., Cariaggi, L.: Investigating neighborhood generation methods for explanations of obscure image classifiers, in *PAKDD (1)*, ser. Lecture Notes in Computer Science, vol. 11439. Springer, pp. 55–68 (2019)
14. Guidotti, R., Monreale, A., Matwin, S., Pedreschi, D.: Black box explanation by learning image exemplars in the latent feature space, in *ECML/PKDD (1)*, ser. Lecture Notes in Computer Science, vol. 11906. Springer, pp. 189–205 (2019)
15. Metta, C., Guidotti, R., Yin, Y., Gallinari, P., Rinzivillo, S.: Exemplars and counterexemplars explanations for skin lesion classifiers, in *Frontiers in Artificial Intelligence and Applications*, vol. 354, (2022)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition, in *CVPR*. IEEE Computer Society, pp. 770–778 (2016)
17. Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Giannotti, F.: Local rule-based explanations of black box decision systems, (2018)
18. Ozer, C., Oksuz, I.: Explainable image quality analysis of chest x-rays, in *MIDL*, vol. 143, pp. 567–580 (2021)
19. Boutorh, A., Rahim, H., Bendoumia, Y.: Explainable ai models for covid-19 diagnosis using ct-scan images and clinical data, *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*, 185–199, (2022)
20. Farahani, F.V., Fiok, K., Lahijanian, B., Karwowski, W., Douglas, P. K.: Explainable ai: a review of applications to neuroimaging data, *Frontiers in Neuroscience*, 16, (2022)
21. Jampani, V., et al.: Assessment of computational visual attention models on medical images, in *ICVGIP*, 80, 1–8 (2012)
22. Yoo, S. H., et al.: Deep learning-based decision-tree classifier for covid-19 diagnosis from chest x-ray imaging, *Frontiers in Medicine*, 7, (2020)
23. Papanastasopoulos, Z., et al.: Explainable ai for medical imaging: deep-learning cnn ensemble for classification of estrogen receptor status from breast mri, in *SPIE*, 11314, (2012)
24. Wang, C., Liu, Y., Wang, F., Zhang, C., Wang, Y., Yuan, M., Yang, G.: Towards reliable and explainable ai model for solid pulmonary nodule diagnosis, arXiv:2204.04219, (2022)
25. Wang, C., Liu, Y., Wang, F., Zhang, C., Wang, Y., Yuan, M., Yang, G.: Explainability of deep neural networks for mri analysis of brain tumors. Int. J. Comput. Assist. Radiol. Surg. **17**, 1673–1683 (2022)
26. Chen, H., Gomez, C., Huang, C.: Explainable medical imaging ai needs human-centered design: guidelines and evidence from a systematic review,. npj Digit. Med. **5**, 156 (2022)
27. Dhurandhar, A., et al.: Explanations based on the missing: towards contrastive explanations with pertinent negatives, *Advances in Neural Information Processing Systems*, 592–603, (2018)
28. Liu, S., Kailkhura, B., Loveland, D., Han, Y.: Generative counterfactual introspection for explainable deep learning, in *IEEE Global Conference on Signal and Information Processing*. IEEE, (2019)
29. Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B., Ghosh, J.: Towards realistic individual recourse and actionable explanations in black-box decision making systems, in *CoRR*. https://arxiv.org/abs/1907.09615, (2019)
30. Samangouei, P., et al.: Model explanation via decision boundary crossing transformations, in *Proceedings of the European Conference on Computer Vision*. ECCV, pp. 666–681 (2018)
31. Singla, S., Pollack, B., Chen, J., Batmanghelich, K.: Explanation by progressive exaggeration, in *ICLR*, (2020)
32. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets, in *NeurIPS Proceedings*, (2014)
33. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
34. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I. J.: Adversarial autoencoders," *CoRR*, https://arxiv.org/abs/1511.05644, (2015)
35. Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., Turini, F.: Factual and counterfactual explanations for black box decision making. IEEE Intell. Syst. **34**(6), 14–23 (2019)
36. Thanh-Tung, H., Tran, T.: Catastrophic forgetting and mode collapse in gans, in *IJCNN*, (2020)
37. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., and, X. C.: Improved techniques for training gans, in *NIPS*, (2016)
38. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P. -A.: Extracting and composing robust features with denoising autoencoders, in *ICML*, (2008)
39. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation, in *ICLR*, (2018)
40. Kruskal, W.H., Wallis, W.A.: Use of ranks in one-criterion variance analysis. Arch Dermatol. **138**, 1562–1566 (2002)
41. Kruger, J., Dunning, D.: Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. Pers. Soc. Psychol. **77**, 1121–1134 (1999)
42. Petsiuk, V., Das, A., Saenko, K.: Rise: randomized input sampling for explanation of black-box models, in *British Machine Vision Conference (BMVC)*, (2018)
43. Kruskal, J.B.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika **29**, 1–27 (1964)
44. Izikson, L., Sober, A.J., Mihm, M.C., Zembowicz, A., et al.: Prevalence of melanoma clinically resembling seborrheic keratosis: analysis of 9204 cases. J. Am. Stat. Assoc. **47**, 583–621 (1952)
45. Ho, T. K.: Random decision forests, in *ICDAR*. IEEE Computer Society, pp. 278–282 (1995)
46. Sondermann, W., Utikal, J.S., Enk, A.H., et al.: Prediction of melanoma evolution in melanocytic nevi via artificial intelligence: a call for prospective data. EJC Euro. J. Cancer **119**, 30–34 (2019)
47. Haggenmüller, S., Maron, R.C., and Helker, A., et al.: Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts," *EJC, Euro. J. Cancer*, 156, (2021)

48. Celebi, M.E., Codella, N., Halpern, A., Shen, D., et at.: Guest editorial skin lesion image analysis for melanoma detection, *IEEE J. Biomed. Health Inform.*, 23(2), (2019)
49. Kawahara, J., Hamarneh, G.: Fully convolutional neural networks to detect clinical dermoscopic features, *IEEE J. Biomed. Health Inform.*, 23(2), (2019)
50. Mahmouei, S. S., Aldeen, M., Stoecker, W. V., and Garnavi, R. et al.: Biologically inspired quadtree color detection in dermoscopy images of melanoma, *IEEE J. Biomed. Health Inform.*, 23(2), (2019)