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 analyze a case study on skin lesion images where we customize an existing XAI approach for explaining a deep learning model able to recognize different types of skin lesions. The explanation is formed by synthetic exemplar and counter-exemplar images of skin lesion and offers the practitioner a way to highlight the crucial traits responsible for the classification decision. A survey conducted with domain experts, beginners and unskilled people proof that the usage of explanations increases the 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 could derive from the intrinsic characteristics of each class and, hopefully, can provide support in the resolution of the most frequent misclassifications 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 consideration in different domains due to their remarkable performance. However, their adoption in sensitive scenarios that involves decision on humans, such as the medical one, has raised ethical concerns about the missing transparency for the decisions based on AI suggestions [1][2]. The need to develop...
AI systems that are able to assist doctors in taking informative decisions, complementing their own knowledge with the information and suggestion yielded by the AI system [3, 4]. However, if the logic for the decisions of AI systems are not available it would be impossible to accomplish 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 clues that help in understanding the reasons for the decision outcome and that do lead to 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 diagnosis 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 [6, 2, 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. First, 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 \( \varepsilon \)-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 issue, in [14] has been proposed ABELE, 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 methodology illustrated in [14] and [3], 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 image classification challenge. We train on the ISIC dataset a state-of-the-art deep learning
classifier using the ResNet CNN architecture, i.e., a ResNet [15]. 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 of the explanations.

The survey has been conducted with domain experts, beginners and unskilled people. The results of the survey show that the usage of explanations increases the 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. 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. Finally, we also design and develop a user interaction module to analyze the explanations returned by ABELE.

The rest of the paper is organized as follows. Section 2 presents the methodology adopted. In Section 3 we illustrate in detail the settings of the case study addressed, and we present the interaction module. Section 4 illustrates the survey and shows the results obtained, while Section 5 presents the analysis of the latent space. Finally, Section 6 summarizes the contribution and proposes future research directions.

2 Methodology

In this section we briefly present the two main components of the methodology adopted to classify and explain the ISIC dataset. Details can be found in [15, 14, 3].

2.1 ResNet Classifier

In order to provide a classifier that performs sufficiently well for downstream learning steps, we choose to train a Neural Network (NN) with an architecture powerful enough to accomplish image classification. In particular, we selected a ResNet, a NN architecture providing validated performance on many complex datasets and tasks [15]. Instead of training the ResNet from scratch, we choose to perform a transfer learning task with a ResNet pretrained 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 NN [16]. To perform the transfer learning, we replace the last fully connected layer with the newly initialized one. The number of output dimension 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.

2.2 ABELE Explainer

As explainer we adopted ABELE (Adversarial black-box Explainer generating Latent Exemplars) [14]. 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 synthetically 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 AAE [17]. The AAE architecture (Fig. 1 left) 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 AAE 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 [18]. The module in Fig. 2 is named llore, as a variant of lore [18]. After the generation process, for any instance $h \in H$, ABELE exploits the disde module (Fig. 1 right) for both checking the validity of $h$ by querying the discriminator and decoding it into $h$. Then, it queries the black-box $b$ with $h$ to get the class $y$, i.e. $b(h) = y$. Given
the local neighborhood $H$, ABELE builds a decision tree classifier $c$ trained on $H$ labeled with $b(H)$. The surrogate tree is intended to locally mimic the behavior of $b$ in the neighborhood $H$. It extracts the decision rule $r$ and counter-factual rules $\Phi$ enabling the generation of exemplars and counter-exemplars. Fig. 2 shows the process that, starting from the image to be explained, leads to the decision tree learning, and to the extraction of the decision and counter-factual rules.

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

### 2.3 Progressive Growing AAE

We summarize here the customization of ABELE we carried on in order to make it usable for the complex image classification task of the ISIC dataset. 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 [19], 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 [20] and Denoising autoencoders [21]. As model of AAE, we adopted a Progressive Growing AAE. Progressive Growing GANs [22] 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 (7×7 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 the discriminator can deal each step with a more structured information. The incremental addition of the layers allows the PGAAE to first learn large scale structure and progressively shift the attention to finer detail.

The PGAAE network is reported in Figure 3. In our implementation we used six blocks of layers in order to have a trained AAE able to reproduce skin lesion images of size 224x224 pixels. In summary, as learning tricks we relayed on minibatch discrimination, denoising and progressive structure. Thanks to the PGAAE, 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.

3 Case Study: Skin Lesion Diagnosis

In this section we illustrate the case study analyzed presenting details for the dataset, and for the training of the classifier, and of the autoencoder.

3.1 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 2 to improve the international effort in melanoma diagnosis 2. 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

2https://challenge2019.isic-archive.com/
The dataset is composed of a training set of 25,331 images of skin lesions and their category (labels); a test set of 8,238 images of which the label is not publicly accessible.

### 3.2 ResNet Training

We separated the training into two parts: 80% samples used for training and 20% for validation. Since the UNK category is a reject option and is not available in the training we focused on the 8 other categories. 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. Resolution of the preprocessed images is $224 \times 224$.
- For the validation and test, each image is firstly rescaled to $256 \times 256$ according to the shorter edge, then cropped at the center into a $224 \times 224$ image.

For the evaluation, we used the same metric adopted in the submission system of ISIC 2019. 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 achieves 0.838 of balanced multi-class accuracy on the test set.

### 3.3 PGAAE Training

A customization of aBeLe was necessary in order to make it usable for the complex image classification task addressed by the ResNet 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 RMSE 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., $224 \times 224$, it is common in 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 variegated and good quality skin lesion images. aBeLe equipped with such PGAAE is able to generate meaningful explanations and it can be tested in a survey involving real participants described in the next section.

### 3.4 ABELE User Interaction Module

We present here the novel interaction module for the explanations returned by aBeLe. The module facilitates the comprehension of both the suggestion

---

3Such preprocessing does not deform the lesions in the image.
Figure 4: User interaction module to present the results of 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.

of the black-box model and the explanation produced by ABELE. Figure 4 shows a screenshot from a web application we implemented to present to the user the outcome of our system on a specific instance. The visual space is organized in two sections. The upper part shows the instance under analysis \( x \) with the classification returned by the ResNet on the left, and a synthetic counter-exemplar image returned by ABELE on the right. In Figure 4 we have a Melanoma under analysis and a quite similar synthetic image classified as Dermatofibroma. Still on the left is reported a list of instances that are similar to the instance under analysis but with a different label. This list, that is the neighborhood generated by ABELE through the PGAAE, gives an indication of the variety of the instances in the latent space around the image analyzed. The counter-exemplar is selected among this list as the image that minimizes a distance similarity with \( x \) and maximizes the classification similarity but w.r.t. a different label. Finally, the lower part of the module shows four exemplars, i.e. a set of images returned by ABELE that have the same label assigned by the ResNet to \( x \).

The ABELE interaction module is implemented in Javascript as a web application. More in detail, it communicates with a backend that exposes the functionalities of the ResNet 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

---

4https://kdd.isti.cnr.it/isic_viz/
has been used for the preparation of the survey presented in the next section.

4 Validation and Survey

We designed a survey with the idea of validating the effectiveness of the explanations returned by ABELE for skin lesion diagnosis. In particular, the main purpose is to validate the effectiveness of the explanations in assisting doctors and medical experts in the diagnosis and treatment of skin cancers, as well as to investigate their confidence in automatic diagnosis models based on black-boxes and on the explanations provided by the explainer.

4.1 Survey Structure

The survey is organized in ten questions composed of various points with the same general structure described in the following. We highlight that, if not specified, for every point a different image $x$ is provided to the user. In the remaining of the treatment we name each question $Q_i$ with $i \in [1, 10]$.

Point 1 ($P_1$). 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 interaction module. In particular we presented to the user two exemplars and two counter exemplars of another lesion class. Participants were asked to classify the given image among two different given classes exploiting the explanation. The idea behind this point is to understand if the explanations returned by ABELE really helps in separating between different images, even for non-expert users. From another perspective, this can be considered as the human evaluation of the usefulness metric synthetically observed in [14].

Point 2 ($P_2$). The participants are presented with a labeled image and they are asked to quantify the level of confidence in the black-box classification (using a 0–100 slider).

Point 3 ($P_3$). The participants are presented with the same labeled image of $P_2$, but this time with the visual aid of the explanation returned by ABELE, and they are asked to quantify his confidence once more after looking at the explanations. The objective of this couple of points is to understand if there is an increase/decrease in the confidence towards the AI after having observed an explanation.

Point 4 ($P_4$). 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 are not informed of the correctness of their prediction nor they received further suggestions by looking back at their previous answers or explanations. In order to investigate recipient’s reaction to an incorrect advice, in $Q_6$ we intentionally entered a wrong classification ($P_2$) followed by further wrong advice concerning exemplars and counter-exemplars ($P_3$). All the other nine instances were correctly presented with the ground truth label.
4.2 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 which 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 a wrong advice from an AI, participants will show a substantial decline in confidence and trust toward that model (implicit assessment through the error inserted).

4.3 Survey Results

A total of 156 participants completed the study. Participants signed up for the survey online, after digitally signed 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 to all the ten questions, we chose to consider only who had completed at least 27% of the questionnaire, corresponding to at least an entire question with an answer for each point. Aggregate demographic statistics of the participants are available in Figure 5. The vast majority of participants have a scientific background, among which a fair number having completed studies in medicine or dermatology.

First, we have attached to each participant a score that measures their performance in P1 that test their ability to classify skin lesion images 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 participants who achieved a score lower than 70%. Different threshold in the range [60% – 90%] were also considered but discarded since they yielded approximately the same statistics results. Although the average performance in itself is remarkable (score 82.02%) even among people who were not domain experts nor had specific medical knowledge (score 78.67%), we are interested in the sub-sample of people specialized in medicine or more specifically in dermatology (score 91.26%). A one-way ANOVA on ranks (Kruskal–Wallis H test) [23] applied to sub-sample A and B shows no significant effects on classification performance given by education level nor 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 is supported at least for domain experts - however also participants from other domains showed a remarkable performance.

Figure 6 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, 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 helps in increasing the model trust. However, Figure 6 suggests that Q3 anomaly was caused by the sub sample of non medical expert. Such an increase of 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%. A plausible reason for this phenomenon could be a substantial decline in participant’s confidence following a wrong suggestion, which is then completely recovered if the first wrong suggestion is followed by other consistent wrong suggestions. Participants reject a wrong advice but tend to adapt and reset their knowledge baseline if the wrong suggestions consistently repeat themselves.

Increase in confidence was not uniform among all participants. The results summarized in Figure 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

![Figure 6: Participants confidence in the classification of the black-box before and after receiving the explanation of ABELE.](image-url)
Figure 7: Participants’ confidence among different age groups (top), education level (center), domains (bottom), before and after explanations.

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 [24]). 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
Figure 8: Participants confidence towards ABELE explanations.

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

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

Finally, in Figure 9 we summarize how exemplars and counter-exemplars helped participants in recognizing lesion classes, as stated by respondents in P4. We note here the same trend observed in analyzing confidence in explanations before and after Q6: both experts and non experts show a notable decline in confidence about exemplars and counter-exemplars helping them in the classification task. As expected, the explanations returned by ABELE are considerably more helpful for medical experts than the general population. Moreover, we observe on the sidelines how exemplars have been more effective than counter-exemplars for both experts and for the general population. It can be argued
that such behavior comes from having considered an 8-class classification task - in a binary classification task, exemplars and counter-exemplars may encompass a similar significance, while, as the number of classes increases, the relative importance of examples tends to grow.

5 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) [25] as a form of non-linear dimensionality reduction to translate information about pairwise distances among latent projections into a configuration of the same cardinality mapped into a 2D Cartesian space. Thus, through MDS we turn the latent space with 256 dimension into a visual space with 2 dimensions. Figure 10 shows the latent encoding of 8 skin cancer classes. It is worth noting that some primary features of benign and malignant cancer can be also retrieved from such 2D projection. Indeed, from Figure 10 it is clear that all skin lesion classes expect 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 [26] 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. Thus, we decided to train a Random Forest (RF) classifier [27] with 500 estimator trees over the 2D MDS space. The RF classifier is able to separate apart Melanoma from Benign Ker-
Melanoma with 85.60% accuracy (see Figure 11-left). Another important class is indeed Melanocytic Nevus that shows a peculiar features that can be partially justified by the odd representation in Figure 10 (2nd plot). 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 [28]. The RF classifier trained over the 2D MDS space is also able to separate Melanoma from Melanocytic Nevus with 78.53% accuracy (see Figure 11-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 [29]. We remark that our ResNet was originally trained on eight classes, while the RF as binary classifiers.

Nowadays, the detection of melanomas is one of the most researched topics in the oncologic domain [30, 31, 32]. It is still extremely hard, both for clinicians and computers, to reliably predict a nevus oncologic transformation into a malignant melanoma. In order to address this evolutionary issue, future research need to take into account the evolution of the same oncologic data during time. Our methods and discovery can hopefully help clinicians (or automatic systems) to precisely assess the likelihood of a benign skin lesion to transform into melanoma.

6 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 really help practitioners. The not trivial and time consuming step is the training of the generative model. ABELE explanations allowed us to perform two analysis. The latent space analysis suggests an interesting repartition of image over the latent space, it can hopefully be helpful in separating apart similar classes of skin le-
sion that are frequently misclassified by humans (benign from malignant). Also, we conducted a survey involving real user experts and not expert of skin cancer and of healthcare domains. The survey supports the hypothesis that explanation methods without a consistent validation are not useful. As future research directions, it will be interesting to apply ABELE explainer to different diseases and health domain, 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 interaction module presented to a real time explanations generator. This improvement would require a lot of efforts and resources as in the current implementation extracting a single explanation can be time consuming depending on the image.

Acknowledgment

This work is partially supported by the European Community H2020 programme under the funding schemes: G.A. 825619 AI4EU (ai4eu.eu), G.A. 952026 HumanE AI Net (humane-ai.eu), G.A. 952215 TAILOR (tailor-network.eu), and the ERC-2018-ADG G.A. 834756 “XAI: Science and technology for the eXplanation of AI decision making”.

References

[1] D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri, and F. Turini, “Meaningful explanations of black box AI decision systems,” in AAAI. AAAI Press, 2019, pp. 9780–9784.

[2] T. Miller, “Explanation in artificial intelligence: Insights from the social sciences,” Artif. Intell., vol. 267, pp. 1–38, 2019.

[3] C. Metta, R. Guidotti, Y. Yin, P. Gallinari, and S. Rinzivillo, “Exemplars and counterexemplars explanations for image classifiers, targeting skin lesion labeling,” in IEEE ISCC, 2021.

[4] C. Panigutti, A. Perotti, and D. Pedreschi, “Doctor XAI: an ontology-based approach to black-box sequential data classification explanations,” in FAT*. ACM, 2020, pp. 629–639.

[5] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, “The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies,” Journal of Biomedical Informatics, vol. 113, 2021.

[6] A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable artificial intelligence (XAI),” IEEE Access, vol. 6, pp. 52138–52160, 2018.

[7] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” ACM Comput. Surv., vol. 51, no. 5, pp. 93:1–93:42, 2019.
[8] M. Sundararajan et al., “Axiomatic attribution for dnn,” in ICML. JMLR, 2017.

[9] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, “Not just a black box: Learning important features through propagating activation differences,” CoRR, vol. abs/1605.01713, 2016.

[10] S. Bach, A. Binder et al., “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” PloS one, vol. 10, no. 7, p. e0130140, 2015.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, ““why should I trust you?”: Explaining the predictions of any classifier,” in KDD. ACM, 2016, pp. 1135–1144.

[12] S. M. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” in NIPS, 2017, pp. 4765–4774.

[13] R. Guidotti, A. Monreale, and L. Cariaggi, “Investigating neighborhood generation methods for explanations of obscure image classifiers,” in PAKDD (1), ser. Lecture Notes in Computer Science, vol. 11439. Springer, 2019, pp. 55–68.

[14] R. Guidotti, A. Monreale, S. Matwin, and D. Pedreschi, “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, 2019, pp. 189–205.

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in CVPR. IEEE Computer Society, 2016, pp. 770–778.

[16] S. J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, 2010.

[17] A. Makhzani, J. Shlens, N. Jaitly, and I. J. Goodfellow, “Adversarial autoencoders,” CoRR, vol. abs/1511.05644, 2015.

[18] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini, “Factual and counterfactual explanations for black box decision making,” IEEE Intell. Syst., vol. 34, no. 6, pp. 14–23, 2019.

[19] H. Thanh-Tung and T. Tran, “Catastrophic forgetting and mode collapse in gans,” in IJCNN, 2020.

[20] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. C. and, “Improved techniques for training gans,” in NIPS, 2016.

[21] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, “Extracting and composing robust features with denoising autoencoders,” in ICML, 2008.

[22] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” in ICLR, 2018.

[23] W. H. Kruskal and W. A. W. and, “Use of ranks in one-criterion variance analysis,” Arch Dermatol., vol. 138, pp. 1562–1566, 2002.
[24] J. Kruger and D. Dunning, “Unskilled and unaware of it: how difficulties in recognizing one’s own incompetence lead to inflated self-assessments,” *Pers. Soc. Psychol.*, vol. 77, pp. 1121–1134, 1999.

[25] J. B. Kruskal, “Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis,” *Psychometrika*, vol. 29, pp. 1–27, 1964.

[26] L. Ezikson, A. J. Sober, and M. C. M. J. et alt., “Prevalence of melanoma clinically resembling seborrheic keratosis: Analysis of 9204 cases,” *Journal of the American Statistical Association*, vol. 47, p. 583–621, 1952.

[27] T. K. Ho, “Random decision forests,” in *ICDAR*. IEEE Computer Society, 1995, pp. 278–282.

[28] W. Sondermann, J. S. Utikal, and A. H. E. et alt., “Prediction of melanoma evolution in melanocytic nevi via artificial intelligence: A call for prospective data,” *EJC European Journal of Cancer*, vol. 119, pp. 30–34, 2019.

[29] S. Haggenmüller, R. C. Maron, and A. H. et alt., “Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts,” *EJC, European Journal of Cancer*, vol. 156, 2021.

[30] M. E. Celebi, N. Codella, A. Halpern, and D. S. et alt., “Guest editorial skin lesion image analysis for melanoma detection,” *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, 2019.

[31] J. Kawahara and G. Hamarneh, “Fully convolutional neural networks to detect clinical dermoscopic features,” *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, 2019.

[32] S. S. Mahmoudi, M. Aldeen, W. V. Stoecker, and R. G. et alt., “Biologically inspired quadtree color detection in dermoscopy images of melanoma,” *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, 2019.