Neurocomputing 630 (2025) 129701

Contents lists available at ScienceDirect

Neurocomputing

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

TIxAI: A Trustworthiness Index for eXplainable AI in skin lesions classification

Cosimo Ieracitano [a,b,*], Francesco Carlo Morabito [a,b], Amir Hussain [c], Muhammad Suffian [d], Nadia Mammone [a]

[a] DICEAM, University Mediterranea of Reggio Calabria, Via Rodolfo Zehender, Loc. Feo di Vito, Reggio Calabria, 89122, Italy
[b] CNR-ISASI, Institute of Applied Sciences and Intelligent Systems "E. Caianiello", Via Campi Flegrei 34, Pozzuoli, Napoli, 80078, Italy
[c] School of Computing, Edinburgh Napier University, Edinburgh, United Kingdom
[d] DIIES, University Mediterranea of Reggio Calabria, Via Rodolfo Zehender, Loc. Feo di Vito, Reggio Calabria, 89122, Italy

ARTICLE INFO

Communicated by R. Yang

Keywords:
eXplainable Artificial Intelligence
Convolutional Neural Network
Trustworthiness
Skin lesion classification

ABSTRACT

Skin cancer is one of the leading causes of mortality worldwide. Early diagnosis can ensure more effective patient treatment and outcomes, but, this is challenging due to the high similarity between different skin lesion types. There is a growing interest in developing Artificial Intelligence (AI)-based systems for automated skin lesion classification. However, current AI models are not transparent, leading to a lack of trust from clinicians who struggle to interpret and validate AI decisions. To this end, in this paper, a fine tuned EfficientNet-B0-based classifier is first developed to classify dermoscopic images of Melanoma (MEL), Nevus (NV) and Seborrheic Keratosis (SK) skin lesions gathered from the International Skin Imaging Collaboration (ISIC) dataset. Next, the explainability of the model is investigated. In particular, a new Trustworthiness Index for eXplainable AI, herein referred to as *TIxAI*, is proposed. The *TIxAI* is based on the difference between the relevance degree of the lesion and non-lesion areas, leading to the conclusion that the higher the *TIxAI*, the more trustworthy the classifier is expected to be. Experimental results support the use of the proposed *TIxAI* to assess and benchmark the reliability of classifiers also in other real-world applications.

1. Introduction

Cancer is a pathological condition characterized by the uncontrolled proliferation of abnormal cells in the body that have the ability to replicate and damage normal body tissues. It is worth noting that according to the International Agency for Research on Cancer (IARC) there are over 10 million cancer-related deaths and more than 20 million new cases worldwide [1]. Among the main classes of cancer, skin cancer is one of the most prevalent and deadly form of cancer worldwide with more than 1.5 million new cases estimated only in 2020. Specifically, melanoma, despite accounts for only less than 5% of skin cancer diagnosis, is responsible for over 65% of skin cancer-related fatalities [2]. Hence, identifying early warning features of melanoma would aid in detecting the disease from the very initial stages, when the disease is more treatable, thereby reducing the likelihood of mortality for the patient. In this context, dermoscopy is a noninvasive imaging technique involving the microscopic examination of the skin surface and allows for the visualization of submacroscopical pigments undetectable by the unaided eye. Such imaging technique aids clinicians in the difficult task of early diagnosing skin cancer [3]. It is worth noting that skin lesions classification depends on several properties such as morphology, shape, color and texture. Visual examination of demoscopic images is a difficult task that strongly depends on the dermatologist's experience, and it also requires considerable time-effort. Moreover, although the existence of well-known diagnostic procedure, such as ABCD (Asymmetrical, Border, Color, Diameter) rule, typically employed by clinicians, visual inspection is not an effective method since it may lead to misdiagnose the actual lesion due to the high visual resemblance among different lesion classes as well as the subjectivity of human interpretation. Hence, there is a great deal of interest in developing automatic tools for dermoscopic image analysis. It is worth noting that conventional computer-aided approaches typically require the segmentation of the lesion area, a hand-crafted features extraction and the final classification [4].

However, due to the increasing deployment of Deep Learning (DL) in several real-world applications [5] and considering the impressive results achieved in the medical field [6,7], several DL-based systems for skin lesion discrimination have been emerging. In this context,

* Corresponding author at: DICEAM, University Mediterranea of Reggio Calabria, Via Rodolfo Zehender, Loc. Feo di Vito, Reggio Calabria, 89122, Italy.
E-mail address: cosimo.ieracitano@unirc.it (C. Ieracitano).

https://doi.org/10.1016/j.neucom.2025.129701
Received 10 October 2024; Received in revised form 16 January 2025; Accepted 8 February 2025
Available online 18 February 2025

most of the existing works are based on the analysis of the International Skin Imaging Collaboration (ISIC) datasets. For example, Li et al. [8] proposed LFN (a Lesion Feature Network based on Convolutional Residual Networks) evaluated on the ISIC-17 dataset consisted of three classes: Melanoma (MEL), Nevus (NV), Seborrheic Keratosis (SK), reporting classification accuracy of 91.2%. Mahbod et al. [9] achieved instead 87.7% using a multiple set of CNNs of different architectures using the same dataset. Gouda W. et al. [10] implemented different models (Convolutional Neural Network (CNN), Resnet50, Inception V3, Inception Resnet) in order to classify dermoscopic images of malignant and benign tumors of ISIC-18 dataset, achieving 85.8% accuracy using in the InceptionV3 network. It is to be noted that ISIC-18 dataset contains seven skin lesions (MEL, NV, Basal cell carcinoma (BCC), Actinic keratosis (AKIEC), Benign keratosis (BKL), Dermatofibroma (DF), Vascular lesion (VASC)), but in [10] a binary classification was addressed. Alwakid G. et al. [11] proposed a modified version of ResNet50 and a custom CNN to classify demoscopic images belonging to seven categories of skin lesions of the HAM10000 dataset [12] (i.e., the training set for the ISIC-18 challenge). The quality of images was improved using the enhanced super-resolution generative adversarial networks and segmentation was also used to segment regions of interest. Experimental results reported accuracy up to 86%. Bassel A. et al. [13] proposed a stacked algorithm, which combined predictions from multiple models (i.e., Resnet50, Xception, VGG16) by using the ISIC-19 dataset [12]. The predictions fed a linear regression model for multi-classification purposes, reporting 90.9% accuracy using the Xception model for feature extraction. Aljohani K. et al. [14] tested different DL models (e.g., DenseNet201, MobileNetV2, ResNet50V2, ResNet152V2, Xception, VGG16, VGG19, GoogleNet) to perform the binary classification, melanoma vs. non-melanoma images, extracted from the ISIC-19, achieving accuracy and F1-score of 73% and 81%, respectively. Alam T.M. et al. [15] proposed an efficient RegNetY-320-based skin cancer classifier for identifying the seven categories of HAM10000 dataset with accuracy 91% by applying augmentation. Meswal H. et al. [16] developed a weighted ensemble system based on InceptionV3, Xception, ResNet50, EfficientNetB4, and MobileNet for binary classification (melanoma vs. nevus), achieving an accuracy of 85.54%; while, recently, Senthil Sivakumar et al. [17] proposed a binary classification based on ResNet50 achieving a maximum accuracy of 94%.

Although the aforementioned systems achieved impressive performance, the interpretability of the model behavior remains vague. In this context, eXplainable Artificial Intelligence (xAI) aims to investigate the *black box* behavior and provide further explanations into the inner working mechanisms that power the AI algorithms [18]. In clinical application, as in melanoma recognition, this holds utmost significance, since a deeper understanding could substantially impact the final decision-making of clinicians [19–22]. In the last years, xAI has been exploited to explain the reasoning behind the specific decisions produced by skin lesions classification systems [23–30]. However, in this context, to the best of our knowledge, there is not any general index to quantify the explainability and consequently the trustworthiness of the developed model. In order to fill this gap, the present study introduces a novel trustworthiness index for explainable AI in skin lesions classification, named *TIxAI*. Specifically, the major contributions of this paper are summarized as follows:

- Development of an automatic multi-class skin lesions classification system based on the fine tuned EfficientNet-B0 and on a revised version of ISIC-17;
- Explainability of the achieved outcomes by means of xAI techniques to evaluate reliability of the proposed model;
- Development of a new index to measure the degree of trustworthiness of xAI in skin lesions classification, here called *TIxAI*;
- Development of a trustworthy skin lesions classification system with potential deployment in clinical setting.

The paper is organized as follows: Section 2 describes the dataset and introduces the proposed methodology, including the developed multi-class skin lesions classifier, the xAI algorithms employed and the definition of the trustworthiness index *TIxAI*; Section 3 reports the achieved experimental results, while Section 4 concludes the paper.

2. Methods

Fig. 1 shows the flowchart of the proposed methodology. It includes the following stages: (i) a *classification stage*, where the classifier is trained to discriminate between skin lesions (MEL/NV/SK); (ii) an *explaining stage*, where an explainability analysis is carried out to evaluate the trustworthiness of the decisions of the trained classifier.

2.1. Skin lesions dermoscopic images dataset preparation

In this study, the publicly available dataset provided by the International Skin Imaging Collaboration (ISIC) as part of the 2017 challenge is used [31]. It is to be noted that, among the datasets provided by ISIC, the 2017 has been selected since it is the most recent dataset that provided both the ground truth mask and the gold standard lesion diagnoses (i.e., the corresponding label) alongside the lesion images. This choice was made as this study focuses on the introduction of an innovative measure of trustworthiness index, which relies on the utilization of the ground truth mask as detailed in the subsequent sections. ISIC-17 consists of 2750 dermoscopic images provided by different international clinical centers and belonging to three categories of skin lesions: “Melanoma” (MEL), “Nevus” (NV), “Seborrheic Keratosis” (SK). Fig. 2 shows examples of dermoscopic images. It is worth noting that the dataset included images with artefacts such as the presence of marking-pen, ruler marks, light reflection, fuzz, dark corners, color charts and air/oil bubbles that may negatively impact the classification process (Fig. 3). For this reason, the dataset was carefully revised by an expert operator removing such images. The final version of ISIC dataset used in this study consisted of 2325 dermoscopic images: 478 related to MEL, 1520 related to NV, 327 related to SK. However, since the dataset resulted highly imbalanced, standard data augmentation techniques (i.e., rotation, flipping and translation) have been applied to MEL and SK classes. In addition, due to variability in size among ISIC images, these were resized to dimension of 224×224 to input the pre-trained models, while keeping the computation time low.

2.2. Skin lesions classification

2.2.1. Customised CNN-based skin lesions classifier

The architecture of the proposed custom CNN is depicted in Fig. 4. It is to be noted that the CNN topology has been selected empirically according to a *trial and error* approach. Several architectures were indeed taken into account, as reported in Table 1. Specifically, the proposed CNN (i.e., CNN_6 of Table 1) consists of: four convolutional layers, each one followed by a normalization layer, a ReLU activation layer and a max pooling layer; five fully connected layers, with 1000, 200, 100, 50 and 10 processing units, respectively. The convolution layers have 32, 64, 128 and 256 filters of size 3×3 for the first convolutional layer and 2×2 for the remaining layers, zero padding and stride equal to 1; while, pooling layers have 2×2 filters with a stride of 2 (with the exception of the fourth layer, which is characterized by a unitary stride). The network ends with a final output classification layer that allows to perform the classification among MEL, NV, SK images. The custom CNN was trained using the Adaptive Moment optimization (ADAM) technique [32] with a low learning rate ($\eta = 0.0001$), an exponential decay rate for the 1st moment estimates ($\beta_1 = 0.9$) an exponential decay rate for the 2nd moment estimates ($\beta_2 = 0.999$) and a small constant for numerical stability ($\epsilon = 1e^{-8}$). It is worth noting

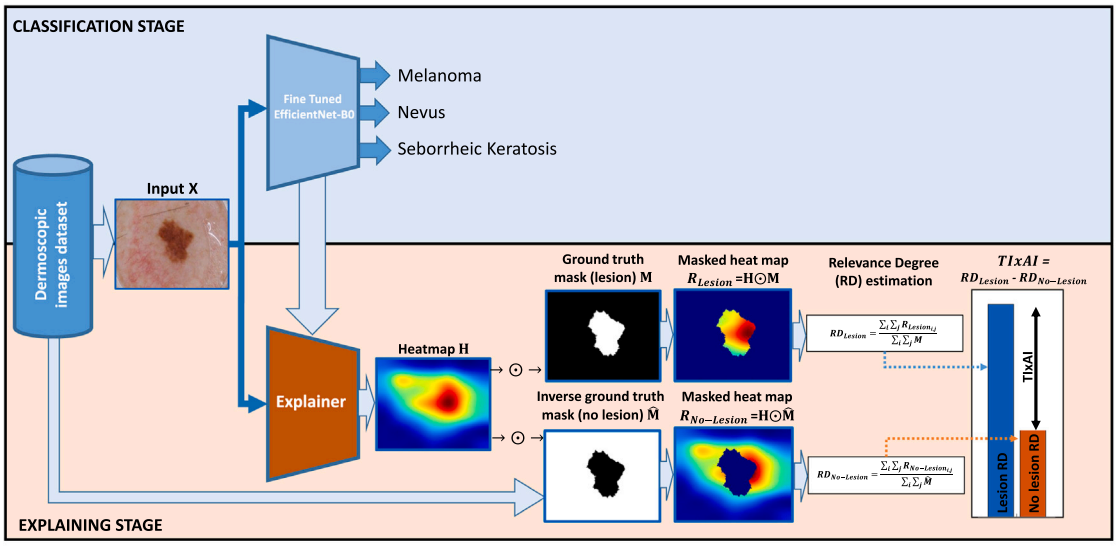


Fig. 1. Flowchart of the proposed methodology consisting of two main steps: classification and explainability. The classification stage includes a fine tuned EfficientNet-B0-based classifier able to discriminate among Melanoma (MEL), Nevus (NV) and Seborrheic Keratosis (SK) lesions. The explaining stage includes an explainer based on xAI techniques (i.e., Grad-CAM, LIME, OSA, SHAP) able to produce the so called *heatmaps*. These are multiplied by the ground truth and the inverse ground truth in order to calculate the relevance of the lesion and non-lesion areas. Finally, the "*Trustworthiness Index for Explainable AI (TIxAI)*" is estimated as the difference of the relevance degree between the lesion and non-lesion area.

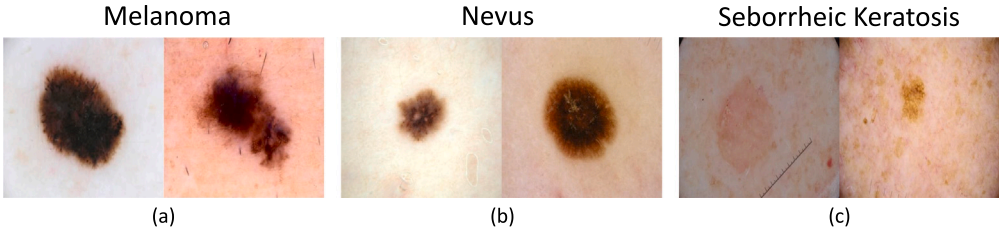


Fig. 2. Examples of Melanoma (MEL), Nevus (NV) and Seborrheic Keratosis (SK) dermoscopic images belonging to the ISIC dataset [31].

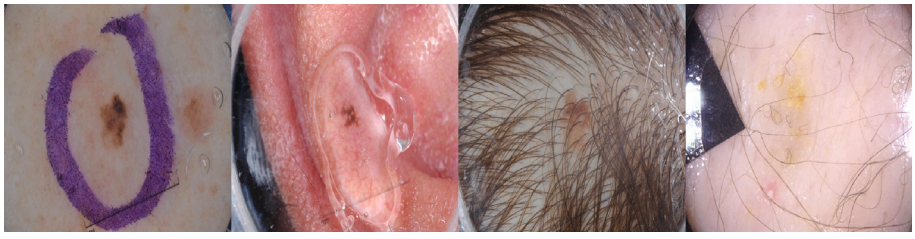


Fig. 3. Examples of dermoscopic images with artefacts.

that since the proposed custom CNN contained a substantial number of learnable parameters (approximately 43.6 million) a maximum of 50 epochs was set. The cross-entropy loss function was monitored during training and the convergence was empirically observed within this range. An early stopping criterion was applied, namely, training was arrested if no improvement was observed in the validation loss for 5 consecutive steps.

2.2.2. Fine tuned CNN-based skin lesions classifiers

Pretrained neural networks allow the transfer of knowledge (i.e. weights) from one task to another. In this study, well-known GoogleNet, ResNet-18, MobileNetv2, EfficientNet-B0, VGG16, initially trained on large ImageNet dataset, are fine-tuned to perform a new classification task. Indeed, the last classification layer is replaced with a new layer with the number of outputs equal to the number of classes of the new dataset, while keeping unchanged the overall architecture of the model. Here, a 3-way output classification layer is used to perform the 3-way skin lesion classification: MEL vs. NV vs. SK. As discussed in Section 3, EfficientNet-B0 outperformed the other models. The architecture of EfficientNet-B0 is depicted in Fig. 5. It is the basic model in the EfficientNet family of CNN architectures, introduced by Tan et al. [33]. Specifically, EfficientNet-B0 starts with a set of processing layers consisting of a standard convolutional layer with 3 × 3 kernel size followed by a batch normalization and a swish activation function. Then, the network is a sequence of Mobile Inverted Bottleneck Convolution (MBConv) modules, i.e., MBConv1 and MBConv6, which consist of a combination of a depth-wise convolution layers, swish activation, and Squeeze-and-Excitation (SE) sub-block capable of enhancing the ability of the model to learn relevant features [34,35]. The EfficientNet-B0 and the other pretrained models were fine-tuned by re-training the

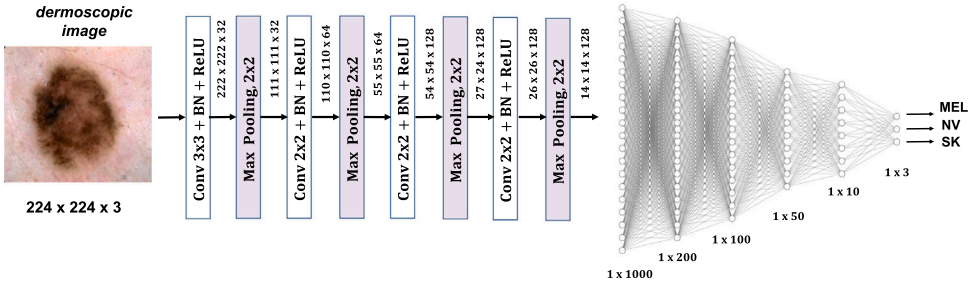


Fig. 4. Architecture of the custom Convolutional Neural Network (CNN), developed to perform the multi-skin lesions classification: MEL vs. NV vs. SK.

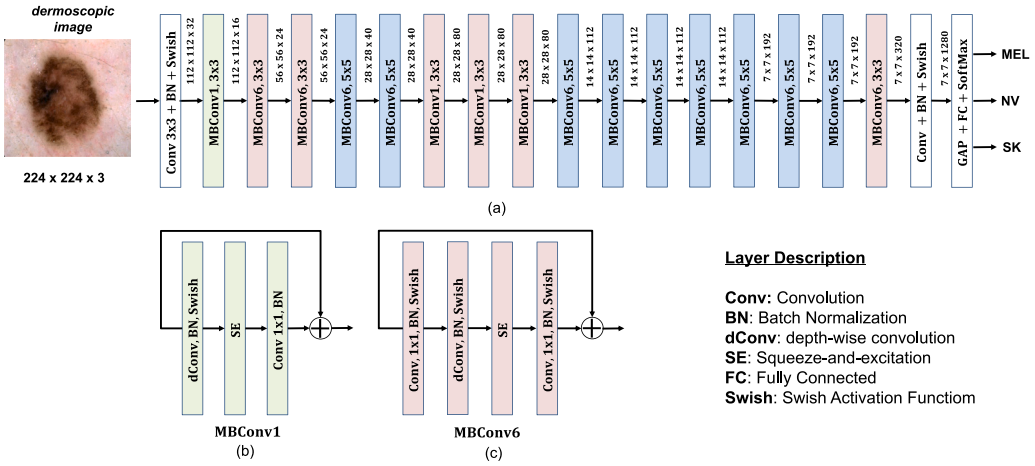


Fig. 5. Architecture of the EfficientNet-B0 used to perform the multi-skin lesions classification: MEL vs. NV vs. SK.

entire architecture on the new dermoscopic image dataset. Training was conducted for a maximum of 10 epochs using ADAM optimizer with the following hyper-parameters: $\eta = 0.0001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e\text{-}8$. The cross-entropy loss function was monitored throughout training and the convergence was empirically observed within 10 epochs. Early stopping was used to stop training when the validation loss showed no improvement for 5 consecutive steps.

Custom and fine-tuned networks were trained and tested in MATLAB R2023b over a workstation with one NVIDIA GeForce RTX 2080 Ti GPU and 64 GB RAM installed. In addition, in order to ensure the reliability of the models, the k-fold cross-validation approach with $k = 7$ was also applied. Hence, results are reported as mean ± standard deviation.

2.3. Explainability of skin lesions classifier

The procedure involves an offline phase of transfer learning using EfficientNet-B0, which is fine-tuned on the dermoscopic images dataset, and a subsequent phase of explainability analysis of the trained classifier's behavior. In this study, the explainability was investigated through Gradient-weighted Class Activation Mapping (Grad-CAM), Local Interpretable Model-Agnostic Explanations (LIME), the Occlusion Sensitivity Analysis (OSA) and the SHapley Additive exPlanations (SHAP) to tentatively explain and interpret the decisions made by the proposed skin lesions classifier. In particular, once the classification phase is completed, given an input image, the classifier is exploited by the *explainer* (i.e., Grad-CAM/LIME/OSA/SHAP) to estimate the relevance to classification of each area in the input image. In other words, the explainer assesses which areas of the input images were more decisive for the network in order to predict the class of the image under analysis. The explainer's output is a *heatmap* (**H**) that associates each area of the input image with a color corresponding to the estimated relevance (red indicating high relevance, blue indicating low relevance). Relevance values are then normalized between zero and one.

2.3.1. Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM is an extension of the CAM method and measures the significance of every neuron in a neural model by analyzing the gradients of the target class as they propagate through the deep network. Specifically, Grad-CAM estimates the gradient of the score of the class c (y^c) with respect to the features maps (F^m), where m is the number of features maps) in a specified convolutional layer. In this study, the features maps of the last convolutional layer are considered. First, the neuron importance weight is computed as follows:

$$\alpha_m^c = \frac{1}{N} \sum_i \sum_j \frac{\partial y^c}{\partial F_{i,j}^m} \tag{1}$$

where N denotes the number of elements in the mth features map F and (i, j) pinpoints the elements. Then, the class-discriminative localization map ($L^c_{Grad-CAM}$) is computed as a weighted combination of F^m with the application of a ReLU:

$$L^c_{Grad-CAM} = ReLU(\sum_m \alpha_m^c F^m) \tag{2}$$

The Grad-CAM output is a *saliency map* where the most relevant input regions are encoded with coloration from blue (low relevance) to red (high relevance) [36]. In this study, Grad-CAM was implemented using the built-in *gradCAM* function available in MATLAB 2023b.

2.3.2. Local interpretable model-agnostic explanations (LIME)

LIME utilizes a perturbation-based algorithm to provide local interpretability by using a surrogate interpretable model. The latter is trained on a newly generated dataset consisting of perturbed instances, weighted around the specific instance being analyzed. This approach minimizes $\mathcal{L}$, metric that quantifies the fidelity of the surrogate model

Table 1
Topology of different custom CNNs for classifying skin lesions images of MEL/NV/SK.

Model	$Conv_1$	$Max\text{-}Pool_1$	$Conv_2$	$Max\text{-}Pool_2$	$Conv_3$	$Max\text{-}Pool_3$	$Conv_4$	$Max\text{-}Pool_4$	HL_1	HL_2	HL_3	HL_4	HL_5	HL_6	Output
CNN_1	filters = 8@ 3 × 3 × 3 s = 1	filters = 2 × 2 $\bar{s}$ = 2	filters = 16@ 2 × 2 × 8 s = 1	filters = 2 × 2 $\bar{s}$ = 2	filters = 32@ 2 × 2 × 16 s = 1	filters = 2 × 2 $\bar{s}$ = 2	–	–	400	–	–	–	–	–	3
CNN_2	filters = 8@ 3 × 3 × 3 s = 1	filters = 2 × 2 $\bar{s}$ = 2	filters = 16@ 2 × 2 × 8 s = 1	filters = 2 × 2 $\bar{s}$ = 2	filters = 32@ 2 × 2 × 16 s = 1	filters = 2 × 2 $\bar{s}$ = 2	–	–	400	200	–	–	–	–	3
CNN_3	filters = 32@ 7 × 7 × 3 s = 1	filters = 2 × 2 $\bar{s}$ = 2	filters = 64@ 4 × 4 × 32 s = 1	filters = 2 × 2 $\bar{s}$ = 2	filters = 128@ 2 × 2 × 64 s = 1	filters = 2 × 2 $\bar{s}$ = 2	–	–	400	200	100	50	25	15	3
CNN_4	filters = 16@ 3 × 3 × 3 s = 1	filters = 2 × 2 $\bar{s}$ = 2	filters = 32@ 2 × 2 × 16 s = 1	filters = 2 × 2 $\bar{s}$ = 2	filters = 64@ 2 × 2 × 32 s = 1	filters = 2 × 2 $\bar{s}$ = 2	filters = 128@ 2 × 2 × 64 s = 1	filters = 2 × 2 $\bar{s}$ = 1	300	150	–	–	–	–	3
CNN_5	filters = 16@ 3 × 3 × 3 s = 1	filters = 2 × 2 $\bar{s}$ = 2	filters = 32@ 2 × 2 × 16 s = 1	filters = 2 × 2 $\bar{s}$ = 2	filters = 64@ 2 × 2 × 32 s = 1	filters = 2 × 2 $\bar{s}$ = 2	filters = 128@ 2 × 2 × 64 s = 1	filters = 2 × 2 $\bar{s}$ = 1	375	75	15	–	–	–	3
CNN_6	filters = 32@ 3 × 3 × 3 s = 1	filters = 2 × 2 $\bar{s}$ = 2	filters = 64@ 2 × 2 × 32 s = 1	filters = 2 × 2 $\bar{s}$ = 2	filters = 128@ 2 × 2 × 64 s = 1	filters = 2 × 2 $\bar{s}$ = 2	filters = 256@ 2 × 2 × 128 s = 1	filters = 2 × 2 $\bar{s}$ = 1	1000	200	100	50	10	–	3

(i.e., g) in approximating the behavior of the complex classifier (i.e., f) locally. The LIME interpretation, denoted as η is derived through the following optimization procedure:

$$\eta(x) = argmin_{g \in G} \ \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{3}$$

where x represents the instance being analyzed, G denotes the set of interpretable models, and $\mathcal{L}(f, g, \pi_x)$ is the fidelity function. This function measures the reliability of the model g (belonging to G) in approximating the complex model f in the local neighborhood defined by π_x. Moreover, $\Omega(g)$ represents the complexity of the interpretable model [37]. In image processing applications, LIME segments the image under analysis into small feature patches known as *superpixels* and identifies the subset of such superpixels that influences the network's decision. Specifically, LIME creates perturbations by selectively including or excluding certain superpixels in the image. In this way, synthetic images are generated. In this study, each pixel within an excluded feature is replaced by the average pixel value of the image. A simpler surrogate interpretable model (here, a linear regression tree) is then used on the perturbed images in order to approximate the classifier's behavior to pinpoint the most important features [37]. In this study, LIME was implemented using the built-in *imageLIME* function available in MATLAB 2023b.

2.3.3. Occlusion sensitivity analysis (OSA)

OSA is a simple technique that identifies which regions of an image are most relevant for a specific classification task. It involves perturbing systematically different patches of the input by overlaying a shifting mask and evaluating the corresponding effect on the network's output. Specifically, the perturbed image (i.e., image with a portion occluded) feeds a pre-trained network (e.g., EfficientNet-B0) evaluating the variation in the classification score. These changes in classification are used to generate a *heatmap* or *saliency map* that use a system of color-coded from blue to red: red indicates higher values and corresponds to the most critical areas contributing to the identification of the specified class, as occluding these regions reduces classification performance. In contrast, blue represents lower values and denotes regions that are less relevant for the task [38]. In this study, a square mask with size of 20% of the input image was applied and moved across the input data with a vertical and horizontal stride of 22. It is to be noted that mask size and stride were determined empirically. OSA was implemented using the built-in *occlusionSensitivity* function available in MATLAB 2023b.

2.3.4. Shapley additive explanations (SHAP)

SHAP [39] is a comprehensive framework for interpreting machine learning (statistical and deep neural network) models grounded in cooperative game theory. It quantifies the contribution of each feature (e.g., a pixel in the case of an image) to the model's output by assigning it a Shapley value, ensuring a fair and consistent distribution of the prediction's attribution among all input features (i.e., pixels). Let $f(x)$ be a machine learning model, the Shapley value ϕ_i for feature i is computed by evaluating the marginal contribution of i across all possible feature subsets:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|N|!}{|S|!(|N| - |S| - 1)!} [f(x_S \cup \{i\}) - f(x_S)] \tag{4}$$

where, N represents the set of all features, S is a subset of features excluding i, x_S is the input restricted to the features in S, and $f(x_S \cup \{i\})$ is the model output when feature i is included. In this study, the Gradient Explainer from the SHAP library was employed to interpret the deep networks under analysis. The Gradient Explainer approximates Shapley values by integrating the gradient of the model output along the path from a baseline input x' to the actual input x. The Shapley value for each feature i is given as:

$$\phi_i = \int_0^1 \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} \cdot (x_i - x_i') \, d\alpha \tag{5}$$

where, $f(x)$ is the model's output for the target class, x' is the baseline (e.g., a black image), and α is the interpolation factor between the baseline and actual input. Gradient Explainer operates locally (for the given input) and approximates feature (pixel) contribution using gradients. Since Gradient Explainer requires a background dataset to estimate feature contributions, in this study, 100 images were used, as background data to compute Shapley values for all deep networks, consistently. It is worth noting that, since Gradient Explainer relies on gradient computations and the propagation of relevance across layers, requiring substantial computational power, memory, and processing resources, SHAP was computed using 100 images as background data to minimize the computational burden for constrained resources. The Gradient Explainer outcome is a saliency map, highlighting pixels in the input image with coloration from blue (low relevance) to red (high relevance). In this study, the Gradient explainer was implemented using *shap.GradientExplainer* and Shapley values, normalized between zero and one, are computed using the built-in function *shap_values* available in *shap* library of PyTorch [40].

Examples of Grad-CAM, LIME, OSA and SHAP application to dermoscopic skin lesions images MEL/NV/SK are reported in Fig. 12.

Trustworthiness index for explainable AI: TIxAI

In the field of xAI, trustworthiness refers to the degree of confidence that users (e.g., medical doctors) can place on outcomes generated by AI systems. In this context, the Trustworthiness Index for eXplainable AI, herein referred to as *TIxAI*, is introduced to provide insights on the transparency of the AI model's decision-making process, with particular reference to skin lesion classification. Specifically, it estimates the ability of the system to provide reliable results by evaluating in what extent the decision depends on the region in the input image that includes the object of interest (i.e., skin lesion area) and not on a region in the input image that contains no relevant, background details (i.e., non-lesion area). Indeed, by quantifying the extent to which the AI model's decisions rely on the region of the skin lesion area (high saliency values associated with the lesion area), the *TIxAI* aims to make the model trustworthy from the clinicians' and patients' perspective by endorsing possible high classification performance through a high trustworthiness degree. Performance is not enough to make a model reliable, as the final user needs to be sufficiently confident that the model's decisions did not depend on unpredictable features related to the non-lesion areas. The proposed *TIxAI* index quantifies the extent to which high saliency regions in the heatmap match with the lesion areas. The presence of high saliency regions in the non lesion area will reduce the *TIxAI* index. The proposed *TIxAI* index is based on the combination of the heatmaps, provided by the xAI techniques, and the ground truth mask corresponding to the image under analysis. In the stage of evaluating a trained model, ground truth masks are therefore necessary. Indeed, for each input image, the ISIC-17 dataset provides also a ground truth mask indicating the "lesion" and "non-lesion" regions based on delineation by experts. Starting from the ground truth mask $\mathbf{M}$, the inverse ground truth mask $\hat{\mathbf{M}}$ is estimated as the complementary. The ground truth mask $\mathbf{M}$ is then overlapped to the heatmap $\mathbf{H}$ by performing the Hadamard product $\mathbf{H} \odot \mathbf{M}$, in this way producing the relevance map of the lesion region $\mathbf{R}_{Lesion}$. The element $\mathbf{R}_{Lesion_{i,j}}$ is zero if the pixel (i, j) does not belong to the lesion region whereas it is equal to the heatmap level of pixel (i, j) if the pixel belongs to the lesion region. The relevance map of the non-lesion region, $\mathbf{R}_{No-Lesion}$ is calculated similarly by overlapping the inverse ground truth mask $\hat{\mathbf{M}}$ to the heatmap $\mathbf{H}$. In the end, the overall relevance degree (*RD*) is obtained by summing the relevance of each pixel (i, j) in the relevance map ($\mathbf{R}_{Lesion}$ or $\mathbf{R}_{No-Lesion}$) and normalizing with respect to the extension of the corresponding region of interest (lesion or non-lesion).

$$RD_{Lesion} = \frac{\sum_i \sum_j \mathbf{R}_{Lesion_{i,j}}}{\sum_i \sum_j \mathbf{M}} \tag{6}$$

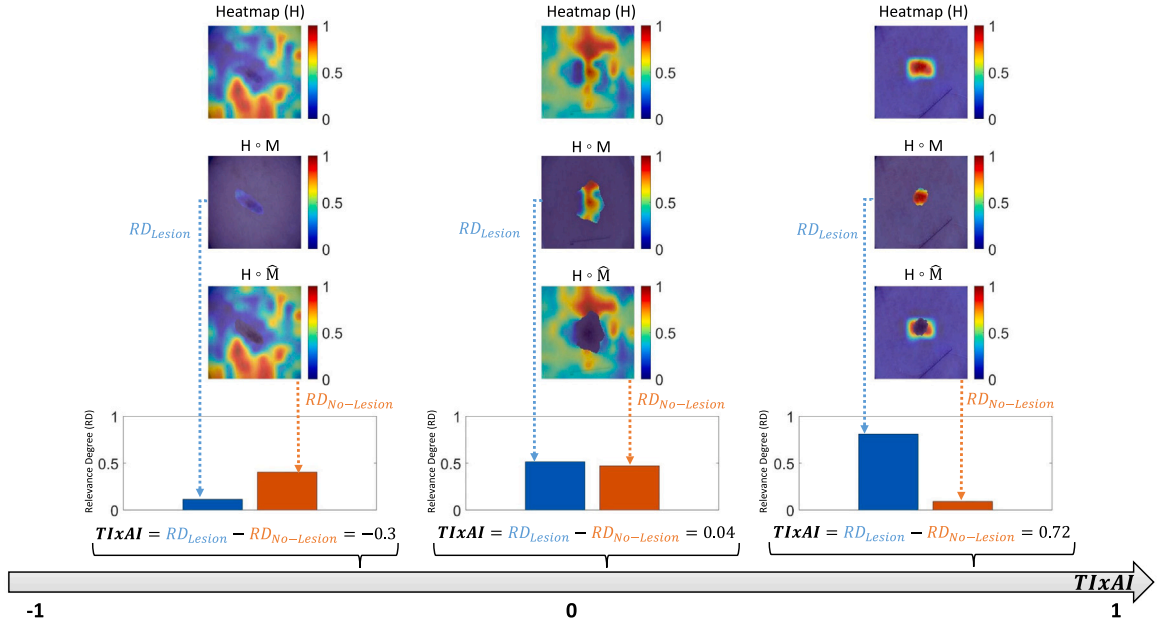


Fig. 6. *TIxAI* is ranged between −1 and 1 (horizontal gray line). The figure shows three representative examples of negative, positive and null values of *TIxAI*. First, the heatmap (**H**) is generated by means of xAI techniques (top row). In this case, the output of the occlusion sensitivity analysis of three dermoscopic images is reported. Then, the Hadamard product $\mathbf{H} \odot \mathbf{M}$ is performed (where **M** is the binary ground truth mask corresponding the dermoscopic image) and the Relevance Degree (RD) of the lesion area ($\mathbf{RD}_{Lesion}$) is estimated according to Eq. (6) (second row). Similarly, the Hadamard product $\mathbf{H} \odot \hat{\mathbf{M}}$ is performed (where $\hat{\mathbf{M}}$ is the inverse ground truth mask), and the Relevance Degree (RD) of the non lesion area ($\mathbf{RD}_{No-Lesion}$) is estimated according to Eq. (7) (third row). Finally, *TIxAI* is computed by estimating the difference between $\mathbf{RD}_{Lesion}$ and $\mathbf{RD}_{No-Lesion}$ (Eq. (8)). In this case, *TIxAI* of −0.3 denotes that the non lesion area is more relevant for the classification; *TIxAI* of 0.04 denotes that the two areas contribute equally; *TIxAI* of 0.72 denotes that the lesion area is more relevant for the classification.

$$RD_{No-Lesion} = \frac{\sum_i \sum_j \mathbf{R}_{No-Lesion_{i,j}}}{\sum_i \sum_j \hat{\mathbf{M}}} \tag{7}$$

$$TIxAI = RD_{Lesion} - RD_{No-Lesion} \tag{8}$$

TIxAI ranges between −1 and 1. Indeed, negative values suggest that the classifier considered more significant the area with no lesion; values close to zero denote that regions with lesion and no-lesion contribute equally to classification; positive values suggest that the lesion area resulted more relevant to classification. Fig. 6 shows three examples of images related to negative (−0.3), null (0.04) and positive (0.72) *TIxAI*. It is reasonable to expect that a classifier is more trustworthy when associated with a large gap between the relevance degree of the lesion area and that of the non-lesion area. This difference is herein referred to as the "Trustworthiness Index for Explainable AI (*TIxAI*)". In conclusion, the higher the *TIxAI*, the more trustworthy the classifier is expected to be.

Classification metrics

The performance of the proposed system is assessed using standard metrics, including Accuracy, Sensitivity, Specificity, Precision, Negative Predicted Value, F1-score [41]. In addition, K-Cohen score is also evaluated [42]. It is defined as:

$$K - Cohen = \frac{Accuracy - Pr(E)}{1 - Pr(E)} \tag{9}$$

with Pr(E) being the probability of random agreement:

$$Pr(E) = \frac{(TP + FP) * (TN + FN)}{(TP + TN + FP + FN)^2} \tag{10}$$

and measures the agreement between the predicted and target classes, with the following score ranges: K-Cohen < 0.01; [0.01–0.20]; [0.21–0.40]; [0.41–0.60]; [0.61–0.80]; [0.81–1.00], indicating, poor agreement, slight agreement, fair agreement, moderate agreement, substantial agreement and very high agreement.

Furthermore, the statistical algorithm known as *t*-distributed stochastic neighbor embedding (*t*-SNE) is used to further analyze the features learned by the model. Specifically, *t*-SNE transforms high-dimensional data points into lower-dimensional feature spaces by translating pairwise distances (using Chebyshev distance in this case) into pairwise joint distributions. It achieves this by minimizing the Kullback–Leibler divergence between the joint probabilities of the low-dimensional and high-dimensional data [43]. In this work, in order to visualize the progress of the learning procedure through the processing modules of EfficientNet-B0, features corresponding to the initial, intermediate and final stage of the network were extracted. In particular, features related to the first MBConv1 (i.e., 112 × 112 × 16), the eighth MBConv6 (i.e., 14 × 14 × 112) and the final global-average-pooling (i.e., 1 × 1 × 1280) were taken into account. The *t*-SNE is thus applied to embed the feature data into a two-dimensional feature space.

3. Results

Classification performance

Table 2 reports the comparative performance of the proposed skin lesions classifiers trained and tested over the revised ISIC-17 dataset (Section 2). In particular, the custom-based classifier yielded the worst outcomes, achieving an average accuracy rate of 64.64 ± 5.67%, sensitivity of 63.69 ± 6.13%, specificity of 78.69 ± 4.42%, precision of 66.90 ± 4.32%, NPV of 78.58 ± 6.86%, F-score of 64.17 ± 5.74% with also a poor agreement score in terms of K-Cohen. In contrast, with the exception of the VGG16 model, which reported an average accuracy of 77%, all the fine tuned networks achieved comparable performance, achieving also substantial and very high agreement of K-Cohen parameter. In particular, as can be noted in Table 2, ResNet-18 and EfficientNet-B0 yielded the highest outcomes with average accuracies of 85.15 ± 1.13% and 87.57 ± 1.32%, respectively. Moreover, it is important to highlight that both networks reported very high agreement of K-Cohen parameter of 0.81 and 0.84, respectively. The

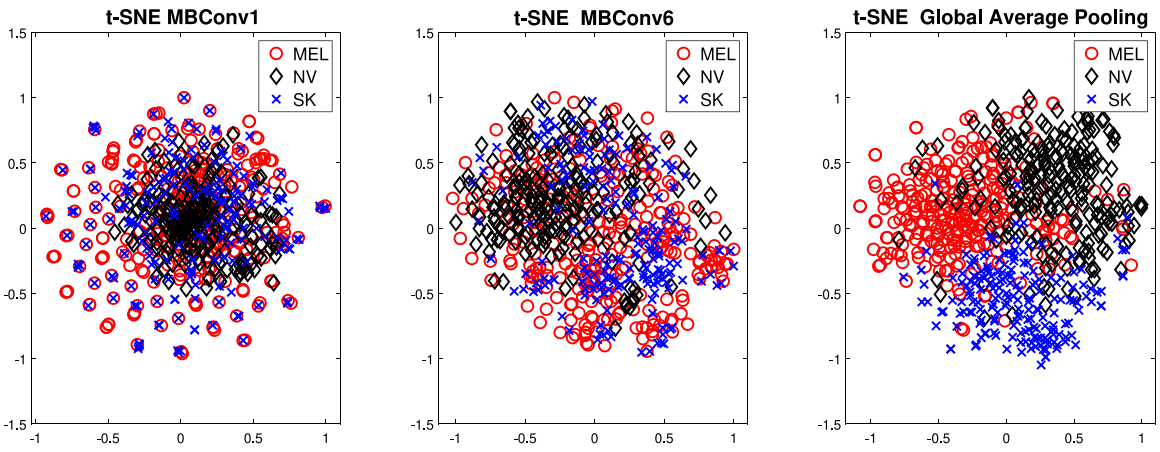


Fig. 7. *t*-SNE visualization of features related to the first MBConv1, the eighth MBConv6 and the final global-average-pooling layer.

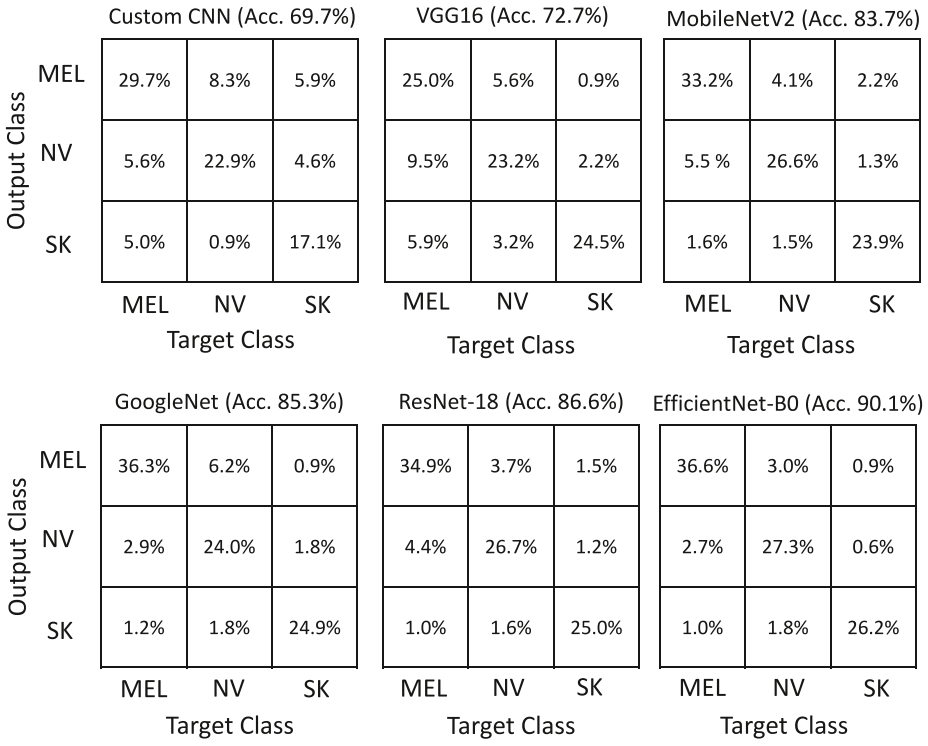


Fig. 8. Confusion matrices for each classification model corresponding to the test fold that yielded the highest accuracy.

confusion matrices of the classification models, corresponding to the test fold that yielded the highest accuracy, are also reported in Fig. 8. As can be observed, the EfficientNet-B0 was able to achieve a maximum classification accuracy of 90.1%. In addition, in order to evaluate the discriminatory capabilities of the proposed network, *t*-SNE was applied to the features extracted by the initial, intermediate, and last layer. Fig. 7 depicts the *t*-SNE scatter plot of MEL (marked in red), NV (marked in black) and SK (marked in blue) class. The projected features are initially overlapped in the earlier stages of the network but became more discriminative after passing through the layers. This observation confirms the deep model proficiency in extracting the most relevant features.

Evaluation of the explainable skin lesions classifier

Fig. 9 shows the average relevance of lesion and non-lesion areas (estimated according to Section 2) achieved by Grad-CAM, LIME, OSA and SHAP. As can be observed, the relevance of lesion areas (blue bars) was greater than the relevance of non-lesion areas (orange bars) for Grad-CAM/LIME/OSA explainer. This means that the proposed skin lesion classifier indeed considered the lesion regions more significant than non-lesion regions. In particular, Grad-CAM explainer reported the maximum value of average relevance degree, namely, 0.5. In contrast, since SHAP is a pixel-level explainability technique in image processing, meaning that the importance of individual pixels in the classification decision tends to be sparse across the image, comparable degree of relevance of lesion and non-lesion area was achieved with SHAP explainer. Fig. 10 shows the boxplots of the proposed *TIxAI* metric (i.e., the difference between the relevance degree of the lesion area and the relevance degree of the non-lesion area) for each explainer. As can be observed, the proposed *TIxAI* had positive median for Grad-CAM, LIME and OSA explainers, confirming the relevant contribution of the lesion areas to the classifier's decision; whereas, *TIxAI* related to SHAP explainer was nearly to zero. This was due to the fact that the relevance degree of the lesion area and the relevance degree of

Table 2
Comparative performance of the proposed custom CNN-based and pre-trained CNN-based skin lesions classifiers.

Model	K-Cohen	Sensitivity [%]	Specificity [%]	Precision [%]	NPV [%]	Fscore [%]	Accuracy [%]	Execution time
Custom CNN	0.55 ± 0.07	63.69 ± 6.13	78.69 ± 4.42	66.90 ± 4.32	78.58 ± 6.86	64.17 ± 5.74	64.64 ± 5.57	10 min
GoogleNet	0.76 ± 0.04	82.26 ± 2.24	89.99 ± 1.65	82.71 ± 2.97	90.05 ± 1.55	81.85 ± 2.97	81.60 ± 2.95	4 min
ResNet-18	0.81 ± 0.01	85.48 ± 1.11	91.96 ± 0.65	85.36 ± 1.00	91.90 ± 0.65	85.37 ± 1.05	85.15 ± 1.13	2 min
MobileNetv2	0.78 ± 0.02	82.82 ± 1.17	90.37 ± 0.74	82.69 ± 1.28	90.34 ± 0.74	82.73 ± 1.21	82.51 ± 1.26	7 min
EfficientNet-B0	0.84 ± 0.02	87.94 ± 1.28	93.34 ± 0.75	87.61 ± 1.30	93.31 ± 0.75	87.73 ± 1.28	87.57 ± 1.32	11 min
VGG16	0.72 ± 0.04	76.83 ± 5.38	87.03 ± 2.65	79.98 ± 2.42	87.74 ± 1.95	76.95 ± 4.82	77.13 ± 4.27	40 min

the non-lesion area were similar, leading to a null value of *TIxAI*. Such result was also confirmed by investigating the correlation between the classification performance of the models and their corresponding *TIxAI* values. Indeed, Fig. 11 shows the bar plot of the *TIxAI* generated by each explainer (i.e., Grad-CAM, LIME, OSA, SHAP) corresponding to the maximum accuracy achieved by every classification model (i.e., Custom CNN, VGG16, MobileNetV2, GoogleNet, ResNet-18, EfficientNet-B0). As can be seen, classifiers with lower classification accuracy (i.e., the Custom CNN and VGG16) showed lower or negative *TIxAI* values. In contrast, classifiers with higher accuracy scores reported higher and positive *TIxAI* values. However, It is to be noted that the *TIxAI* estimation based on SHAP explainer was nearly null for all classifiers, which is reflected in a bar positioned close to zero in the figure. Fig. 12 depicts sample instances of dermoscopic images belonging to MEL, SK and NV class, along with the heatmaps that highlight the areas in the input images that resulted more relevant to classification as MEL, SK or NV (according to Grad-CAM/LIME/OSA/SHAP) and along with the estimated *TIxAI* value. Positive *TIxAI* scores were recorded in each example for Grad-CAM/LIME/OSA, with a maximum value of *TIxAI*=0.79 for the MEL image and by means LIME explainer. Low *TIxAI* was instead observed especially in SK images, with *TIxAI* of 0.43 and 0.58; while, negative and very low *TIxAI* scores (roughly zero) were achieved using SHAP. Finally, Fig. 13 shows the scatter plots of the *TIxAI* scores achieved by using LIME vs. OSA, LIME vs. Grad-CAM and OSA vs. Grad-CAM for each test image rightly classified as MEL, NV, SK, denoted as a black dot in the Figure. It is to be noted that since the *TIxAI* for SHAP explainer was approximately null for each instance, as clearly visible in Figs. 10 and 11, the scatter plot and the statistical comparison of SHAP vs. LIME, SHAP vs. OSA and SHAP vs. GRAD-CAM were not informative and were not taken into account in Fig. 13. Dots positioned above the diagonal line indicate those samples for which one explainer worked better than another. In particular, as can be observed from the distributions of Fig. 13, Grad-CAM outperformed the other techniques. In addition, the statistical significance of the difference between the *TIxAI* of each comparison (e.g., LIME vs. Grad-CAM) was also estimated using the Wilcoxon Rank-Sum Test. The null hypothesis, which suggests that the two sets of samples, namely, the *TIxAI* scores associated with one explainer (such as LIME) and those linked with the second explainer (such as Grad-CAM), are independent samples drawn from identical continuous distributions with equal medians, was evaluated. A p-value <0.05 was achieved in each comparison, resulting in the rejection of the null hypothesis.

Comparison with the state-of-the-art

The proposed study was compared with other works that investigated the application of explainable AI to skin lesions classifiers trained and tested on ISIC datasets. For example, Yang et al. [23] employed the Class Activation Mapping (CAM) to highlight the discriminating image areas used by the developed ResNet50 based-classifier to identify the class under analysis. Wang et al. [24] proposed an interoperable multi-modal CNN for skin lesion detection based on Grad-CAM algorithm. This technique was also exploited by Zia et al. [27] and Ahmad et al. [28] to investigate the explainability of the proposed binary classifier (i.e., the modified-DenseNet201) and multi-class classifier (i.e. Xception-ShuffleNet), respectively. Singh et al. [26] used XRAI,

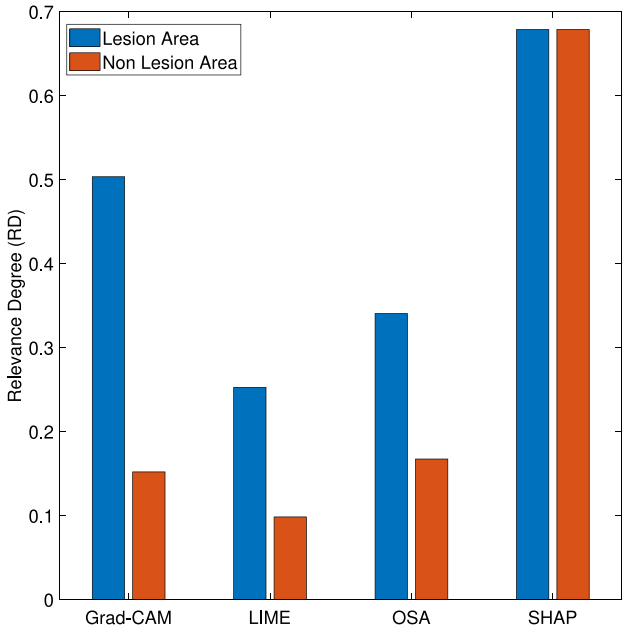


Fig. 9. Average relevance degree of lesion and non-lesion areas, estimated by each explainer (i.e., Grad-CAM, LIME, OSA, SHAP).

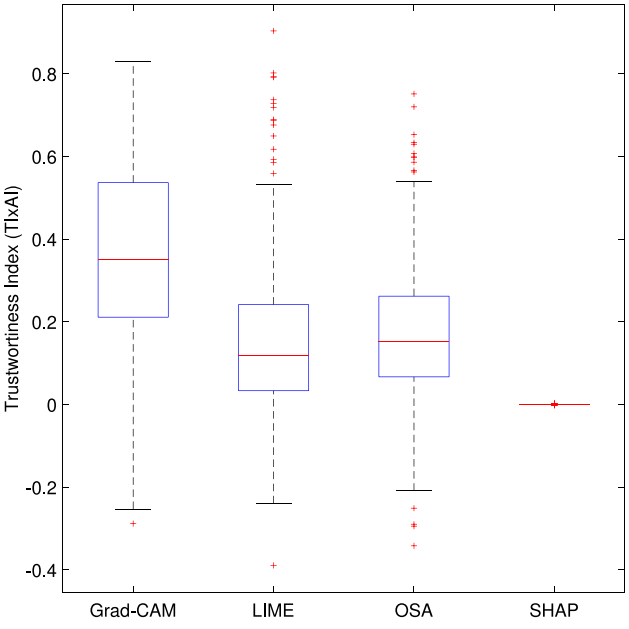


Fig. 10. Boxplot analysis of the proposed *TIxAI* metric estimated for each explainer (i.e., Grad-CAM, LIME, OSA, SHAP).

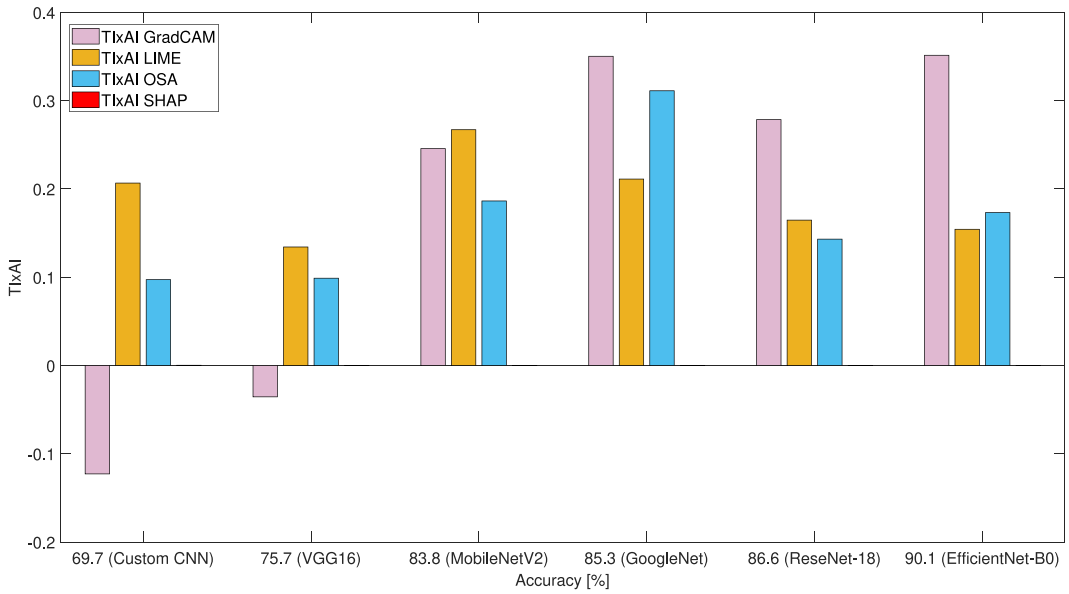


Fig. 11. Bar plot representation of the *TIxAI* generated by each explainer (i.e., Grad-CAM, LIME, OSA, SHAP) compared to the maximum accuracy score achieved by every classification model (i.e., Custom CNN, VGG16, MobileNetV2, GoogleNet, ResNet-18, EfficientNet-B0). It is worth noting that the *TIxAI* estimation based on SHAP explainer was nearly null, which is reflected in a bar positioned close to zero in the figure.

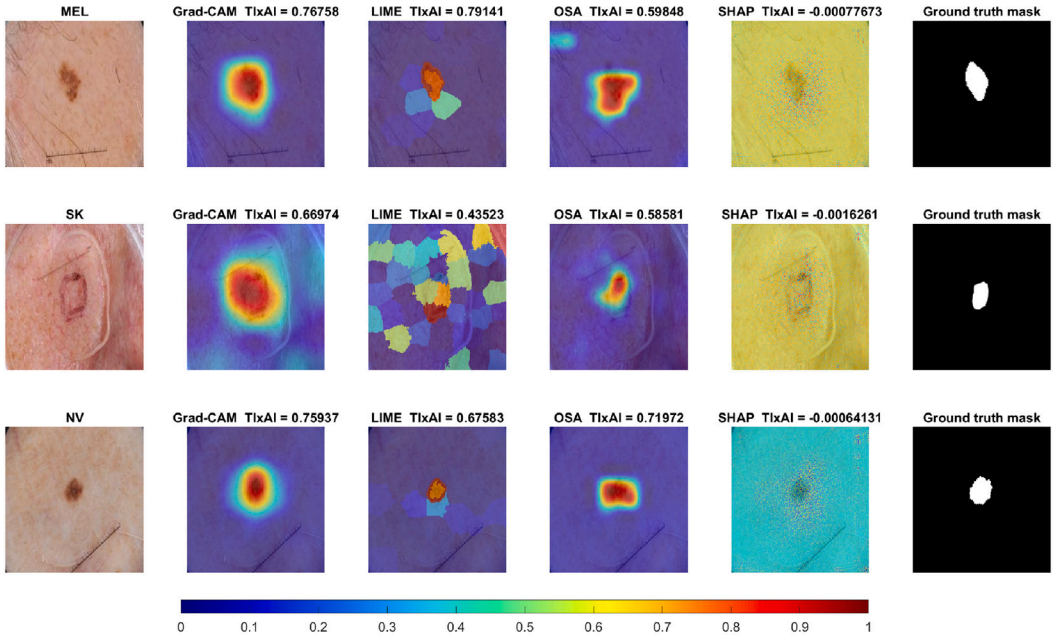


Fig. 12. Heatmaps produced by Grad-CAM, LIME, OSA and SHAP based-explainers for three sample dermoscopic images, belonging to MEL, SK and NV classes, that were correctly classified by the fine-tuned EfficientNet-B0 classifier. It is to be noted that Grad-CAM, LIME and OSA identify the relevant input areas contributing to MEL/SK/NV classification, whereas SHAP, as a pixel-level explainability technique in image processing, highlights the significance of individual pixels, resulting in only a few relevant pixels across the image. In the figure, the importance of the input region or pixel was emphasized using a color gradient ranging from blue (indicating low significance) to red (indicating high significance). The figure reports also the corresponding *TIxAI* index estimated comparing the overall relevance degree of lesion and non-lesion areas. Lesion and non lesion areas were derived from the ground truth mask. Positive *TIxAI* values indicate that the relevance of the lesion region is higher than the relevance of the non-lesion region in the classification process (vice-versa for negative values).

Grad-CAM and Guided Backprop to SkiNet, a skin lesion diagnosis DL network developed to classify the seven categories of ISIC-18 dataset. Nigar et al. [25] applied LIME to explain the proposed ResNet-18-based framework capable of achieving accuracy rate up to 94.47% in the multi-skin lesions classification. Recently, Supriyanto et al. [29] used various pre-trained CNN models and argumentation techniques for classifying skin lesions of HAM10000 dataset, reporting 96.90% accuracy. Moreover, SHAP was also analyzed in an attempt to interpret the achieved performance. Finally, Ahmad et al. [28] proposed an explainable skin lesion classification system based on an optimized CNN, incorporating Grad-CAM and Grad-CAM++ to explain the model's decisions. It is worth noting that all the aforementioned studies provided only qualitative visual explanations of the achieved results without quantifying the overall reliability of the model. In contrast, in this study, xAI was applied to the developed classifier and *TIxAI*, a novel metric to quantify the degree of trustworthiness of the explained classifier, was also introduced.

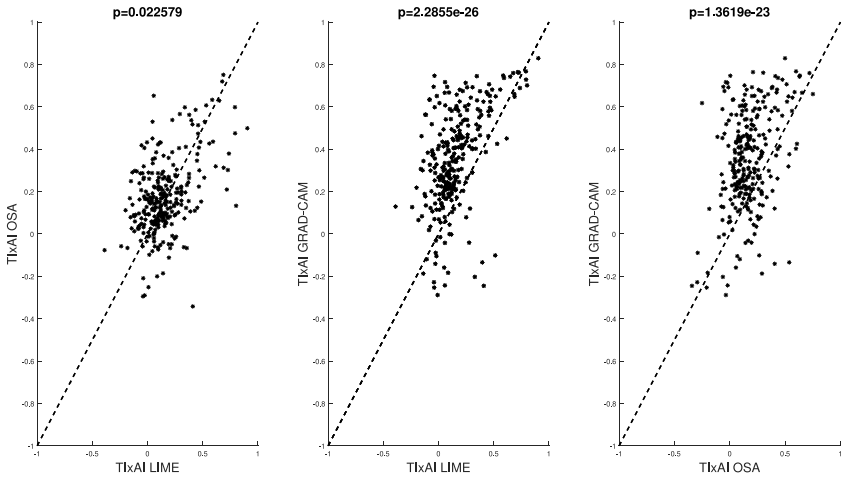


Fig. 13. Scatter-plots of the *TIxAI* estimated by Grad-CAM, LIME and OSA explainer. Dots above the diagonal indicate the samples for which one explainer (e.g., Grad-CAM) outperformed the other (e.g., LIME). The significance of these improvements was assessed using the Wilcoxon Rank-Sum Test ($p < 0.05$).

4. Conclusion

The primary objective of this paper was to investigate the trustworthiness of the proposed skin lesions classifier in order to improve the model reliability. For this purpose, the public "International Skin Imaging Collaboration" dataset, which comprises various versions (ISIC-16, ISIC-17, ISIC-18, HAM10000, ISIC-19, ISIC-20), was considered. Given the specific objective of introducing a novel xAI metric that exploits ground truth response masks alongside class labels, the most recent dataset, i.e. ISIC-17, which provided both these types of both information, was selected for further analysis and experimentation. The proposed methodology comprised a classification and explaining stage. The classification stage included a fine tuned EfficientNet-B0-based classifier that was able to achieve high classification performance (i.e., accuracy of 87.57 ± 1.32%). The explaining stage involved explainability analysis of the proposed classifier using state-of-the-art xAI techniques, namely, Grad-CAM, LIME, OSA and SHAP. In addition, in order to quantitatively assess the reliability of the classifier, a novel metric, referred to as the "Trustworthiness Index for Explainable AI (*TIxAI*)" was introduced to measure the degree of trustworthiness of the classifier, based on the heatmaps produced by the explainers and on the ground truth masks.

To the best of our knowledge, this is the first work that introduces a measure of trustworthiness of explainability analysis in skin lesions classification. Experimental results reported positive average values of *TIxAI*, indicating that the relevance degree of the lesion area was higher than the non-lesion area, in particular for Grad-CAM, LIME, OSA explainer. *TIxAI* values of zeros were instead observed with SHAP explainer. This is due to the fact that SHAP does not inherently account for spatial relationships in image data. Lesion areas in images are contiguous regions, whereas SHAP treats each pixel (feature) independently during its calculations, leading to diluted or uniform attributions for lesion and non-lesion regions. In general, for the classifier to be considered trustworthy, in addition to pursuing high classification performance, the authors argue that a positive *TIxAI* is necessary, indicating that the relevance of the lesion region is higher than the relevance of the non-lesion region. Further, if the classifier recognizes the image correctly but with a negative *TIxAI*, this indicates that the network considered the region with no lesion as more relevant, suggesting that other factors may have influenced the classifier's decision. Since enhanced trustworthiness of AI models is strictly necessary for the case of healthcare applications, the present work aims to make a contribution in the quantitative assessment of models' reliability.

The proposed TIxAI can be considered a general evaluation metric with potential applications beyond skin cancer diagnosis. It can be applied, in principle, in a wide range of medical scenarios [44], as well as other contexts [45]. In medical imaging domains, including radiography (X-rays), neuroimaging, magnetic resonance imaging (MRI), computed tomography (CT), and hematological imaging (e.g., red blood cell analysis), *TIxAI* can assess the trustworthiness of xAI models by evaluating the degree of relevance of specific regions of interest, such as tumors, lesions, or other anatomical features. For example, in MRI-based brain tumor diagnosis, *TIxAI* could validate whether the xAI technique, used to evaluate the reliability of the model, effectively focuses on the tumor boundaries, ensuring that these regions are priority considerations in the decision-making process. For the case of X-rays based lung nodule detection, *TIxAI* could verify whether the AI detection system correctly focuses on the nodule or mass and not on irrelevant areas. In CT-based colorectal cancer screening, *TIxAI* could assess whether the AI and xAI model correctly focuses on the polyp or tumor region, ensuring it does not mistakenly identify surrounding healthy colon tissues. *TIxAI* can also enhance the trustworthiness of AI systems in other domains [46,47]. For instance, in autonomous driving, *TIxAI* could ensure that AI models accurately detect, for example, pedestrians or obstacles, enabling safe and effective navigation. In object recognition tasks, such as ships detection in satellite imagery, *TIxAI* could help validate whether the AI focuses on relevant patterns like full shapes of ships or wakes. In nanotechnology and manufacturing quality control applications, such as anomaly detection in scanning electron microscope (SEM) images of electrospun nanofibers, *TIxAI* could evaluate whether AI models effectively identify areas of defects, such as beads.

Although *TIxAI* has a wide range of potential applications both in clinical and non-clinical contexts, the main limitation of *TIxAI* is its dependence on annotated masks. Whilst annotated masks are essential to the definition of the *TIxAI* metric, their creation requires significant time and effort from domain-specific expertise. For example, annotating medical scans requires manual input from trained radiologists. This dependency limits the applicability of *TIxAI* in scenarios where annotated masks are unavailable. However, it is worth noting that annotated masks are essential in image segmentation for object detection as they represent the gold standard for evaluating segmentation accuracy. In the future, automatic segmentation tools based on the collaboration of multiple centers sharing data and annotations will be included in the proposed framework to assist in automating the definition of masks. This tool could significantly speed up the segmentation process and reduce the burden on experts, making the process more efficient. However, experts validation will still be indispensable as their assessment cannot be disregarded, especially in clinical contexts. The role of experts in reviewing and validating annotations ensures the

segmentation is precise and relevant. Automated methods might not always capture the subtleties that an experienced operator can detect, especially in complex clinical situations. The expert role in verifying and, where required, correcting the automated masks ensures that the final segmentation remains accurate and reliable for clinical decision-making. In addition, the proposed *TIxAI* depends on the output of the explainer. The better the explainer the more reliable *TIxAI* will be. In the future, novel explaining approaches will be proposed. Finally, motivated by the promising results of the TIxAI score, the new metric will also be employed to analyze the explainability of state-of-the-art classifiers to benchmark their performance on other real-world classification tasks.

CRediT authorship contribution statement

Cosimo Ieracitano: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Francesco Carlo Morabito:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Conceptualization. **Amir Hussain:** Writing – review & editing, Writing – original draft, Validation. **Muhammad Suffian:** Software, Validation, Writing – review & editing. **Nadia Mammone:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Conceptualization.

Declaration of competing interest

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

Acknowledgments

This work was funded in part by the European Union - Next Generation EU - PRIN 2022 program, Italian Ministry of University and Research (MUR), project title: "EXEGETE: Explainable Generative Deep Learning Methods for Medical Signal and Image Processing" (project code: 2022ENK9LS, CUP: C53D23003650001); in part by the Programma Operativo Nazionale (PON) "Ricerca e Innovazione" 2014–2020 CCI2014IT16M2OP005 (CUP C35F21001220009 code: I05); in part by the POS RADIOAMICA project funded by the Italian Minister of Health (CUP: C33C22000380006); in part by the Fa.Per.M.E. project funded by the Italian Minister of Health (project code: T3-AN-15, CUP: C33C22000390006); in part by the Next Generation EU - Italian NRRP, Mission 4, Component 2, Investment 1.5, call for the creation and strengthening of 'Innovation Ecosystems', building 'Territorial R&D Leaders' (Directorial Decree n. 2021/3277) - project Tech4You - Technologies for climate change adaptation and quality of life improvement, n. ECS00000009 (CUP: C33C22000290006). This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them. Prof. A. Hussain acknowledges the support of the UK Engineering and Physical Sciences Research Council (EPSRC) Grants Ref. EP/T021063/1(COG-MHEAR) and EP/T024917/1 (NATGEN). The authors wish to thank Mr. Marco Renato and Mrs Chiara Latella, Msc students, for their enthusiastic interest in the very early stages of the project.

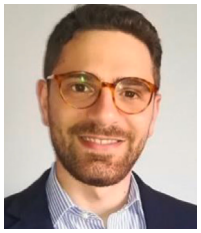
Data availability

This study uses the publicly available dataset provided by the International Skin Imaging Collaboration (ISIC) as part of the 2017 challenge [31].

References

[1] WHO report on cancer: setting priorities, investing wisely and providing care for all, 2024, https://www.who.int/publications/i/item/9789240001299. (Accessed: 17 May 2024).
[2] International agency for reseach on cancer, 2024, https://www.iarc.who.int/cancer-type/skin-cancer/. (Accessed: 24 January 2024).
[3] E. Errichetti, G. Stinco, Dermoscopy in general dermatology: a practical overview, Dermatol. Ther. 6 (2016) 471–507.
[4] A. Maiti, B. Chatterjee, A.S. Ashour, N. Dey, Computer-aided diagnosis of melanoma: a review of existing knowledge and strategies, Curr. Med. Imaging 16 (7) (2020) 835–854.
[5] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, F.E. Alsaadi, A survey of deep neural network architectures and their applications, Neurocomputing 234 (2017) 11–26.
[6] N. Mammone, C. Ieracitano, R. Spataro, C. Guger, W. Cho, F.C. Morabito, A few-shot transfer learning approach for motion intention decoding from electroencephalographic signals, Int. J. Neural Syst. (2023) 2350068.
[7] N. Mammone, C. Ieracitano, H. Adeli, F.C. Morabito, AutoEncoder filter bank common spatial patterns to decode motor imagery from EEG, IEEE J. Biomed. Heal. Informatics (2023).
[8] Y. Li, L. Shen, Skin lesion analysis towards melanoma detection using deep learning network, Sensors 18 (2) (2018) 556.
[9] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot, C. Wang, Fusing fine-tuned deep features for skin lesion classification, Comput. Med. Imaging Graph. 71 (2019) 19–29.
[10] W. Gouda, N.U. Sama, G. Al-Waakid, M. Humayun, N.Z. Jhanjhi, Detection of skin cancer based on skin lesion images using deep learning, in: Healthcare, MDPI, 2022, p. 1183.
[11] G. Alwakid, W. Gouda, M. Humayun, N.U. Sama, Melanoma detection using deep learning-based classifications, in: Healthcare, MDPI, 2022, p. 2481.
[12] M. Naqvi, S.Q. Gilani, T. Syed, O. Marques, H.-C. Kim, Skin cancer detection using deep learning—A review, Diagnostics 13 (11) (2023) 1911.
[13] A. Bassel, A.B. Abdulkareem, Z.A.A. Alyasseri, N.S. Sani, H.J. Mohammed, Automatic malignant and benign skin cancer classification using a hybrid deep learning approach, Diagnostics 12 (10) (2022) 2472.
[14] K. Aljohani, T. Turki, Automatic classification of melanoma skin cancer with deep convolutional neural networks, Ai 3 (2) (2022) 512–525.
[15] T.M. Alam, K. Shaukat, W.A. Khan, I.A. Hameed, L.A. Almuqren, M.A. Raza, M. Aslam, S. Luo, An efficient deep learning-based skin cancer classifier for an imbalanced dataset, Diagnostics 12 (9) (2022) 2115.
[16] H. Meswal, D. Kumar, A. Gupta, S. Roy, A weighted ensemble transfer learning approach for melanoma classification from skin lesion images, Multimedia Tools Appl. (2023) 1–23.
[17] M.S. Sivakumar, L.M. Leo, T. Gurumekala, V. Sindhu, A.S. Priyadharshini, Deep learning in skin lesion analysis for malignant melanoma cancer identification, Multimedia Tools Appl. 83 (6) (2024) 17833–17853.
[18] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.
[19] F.C. Morabito, C. Ieracitano, N. Mammone, An explainable artificial intelligence approach to study MCI to AD conversion via HD-EEG processing, Clin. EEG Neurosci. 54 (1) (2023) 51–60.
[20] C. Ieracitano, N. Mammone, A. Hussain, F.C. Morabito, A novel explainable machine learning approach for EEG-based brain-computer interface systems, Neural Comput. Appl. 34 (14) (2022) 11347–11360.
[21] V. Hassija, V. Chamola, A. Mahapatra, et al., Interpreting black-box models: A review on explainable artificial intelligence, Cogn. Comput. 16 (2024) 45–74.
[22] I. Arshad Choudhry, S. Iqbal, M. Alhussein, K. Aurangzeb, A.N. Qureshi, A. Hussain, A novel interpretable graph convolutional neural network for multimodal brain tumor segmentation, Cogn. Comput. 17 (1) (2025) 1–25.
[23] J. Yang, F. Xie, H. Fan, Z. Jiang, J. Liu, Classification for dermoscopy images using convolutional neural networks based on region average pooling, IEEE Access 6 (2018) 65130–65138.
[24] S. Wang, Y. Yin, D. Wang, Y. Wang, Y. Jin, Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis, IEEE Trans. Cybern. 52 (12) (2021) 12623–12637.
[25] N. Nigar, M. Umar, M.K. Shahzad, S. Islam, D. Abalo, A deep learning approach based on explainable artificial intelligence for skin lesion classification, IEEE Access 10 (2022) 113715–113725.
[26] R.K. Singh, R. Gorantla, S.G.R. Allada, P. Narra, SkiNet: A deep learning framework for skin lesion diagnosis with uncertainty estimation and explainability, Plos One 17 (10) (2022) e0276836.
[27] M. Zia Ur Rehman, F. Ahmed, S.A. Alsuhibany, S.S. Jamal, M. Zulfiqar Ali, J. Ahmad, Classification of skin cancer lesions using explainable deep learning, Sensors 22 (18) (2022) 6915.

[28] N. Ahmad, J.H. Shah, M.A. Khan, J. Baili, G.J. Ansari, U. Tariq, Y.J. Kim, J.-H. Cha, A novel framework of multiclass skin lesion recognition from dermoscopic images using deep learning and explainable AI, Front. Oncol. 13 (2023) 1151257.
[29] C. Supriyanto, A. Salam, J. Zeniarja, A. Wijaya, Two-stage input-space image augmentation and interpretable technique for accurate and explainable skin cancer diagnosis, Computation 11 (12) (2023) 246.
[30] K. Mridha, M.M. Uddin, J. Shin, S. Khadka, M. Mridha, An interpretable skin cancer classification using optimized convolutional neural network for a smart healthcare system, IEEE Access (2023).
[31] N.C. Codella, D. Gutman, M.E. Celebi, B. Helba, M.A. Marchetti, S.W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al., Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic), in: 2018 IEEE 15th International Symposium on Biomedical Imaging, ISBI 2018, IEEE, 2018, pp. 168–172.
[32] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.
[33] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.
[34] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[35] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.
[36] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[37] M.T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you? Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[38] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, Springer, 2014, pp. 818–833.
[39] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS '17, Curran Associates Inc., Red Hook, NY, USA, ISBN: 9781510860964, 2017, pp. 4768–4777.
[40] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035, URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.
[41] N. Japkowicz, M. Shah, Performance evaluation in machine learning, Mach. Learn. Radiat. Oncology: Theory Appl. (2015) 41–56.
[42] M.L. McHugh, Interrater reliability: the kappa statistic, Biochem. Medica 22 (3) (2012) 276–282.
[43] G.E. Hinton, S. Roweis, Stochastic neighbor embedding, Adv. Neural Inf. Process. Syst. 15 (2002).
[44] B.H. Van der Velden, H.J. Kuijf, K.G. Gilhuijs, M.A. Viergever, Explainable artificial intelligence (XAI) in deep learning-based medical image analysis, Med. Image Anal. 79 (2022) 102470.
[45] V. Terziyan, O. Vitko, Explainable AI for industry 4.0: semantic representation of deep learning models, Procedia Comput. Sci. 200 (2022) 216–226.
[46] N. Zeng, H. Li, Z. Wang, W. Liu, S. Liu, F.E. Alsaadi, X. Liu, Deep-reinforcement-learning-based images segmentation for quantitative analysis of gold immunochromatographic strip, Neurocomputing 425 (2021) 173–180.
[47] X. Chen, R. Yang, Y. Xue, C. Yang, B. Song, M. Zhong, A novel momentum prototypical neural network to cross-domain fault diagnosis for rotating machinery subject to cold-start, Neurocomputing 555 (2023) 126656.



Cosimo Ieracitano received the M.Eng. (summa cum laude) and Ph.D. (with additional label of Doctor Europaeus) degrees from the University Mediterranea of Reggio Calabria (UNIRC), Italy, in 2013 and 2019, respectively. He is currently working as an Assistant Professor at the DICEAM Department of the same University (UNIRC). Formerly, he was a Research Fellow at the University Mediterranea of Reggio Calabria, a Visiting Ph.D. Fellow at the University of Stirling (UK) and a Visiting Master Student at ETH Zurich (Switzerland). He has co-authored one international patent and has publications in peer reviewed international journals/conference proceedings in several fields of engineering, such as: artificial intelligence, biomedical signal processing, brain computer interface, information theory and material informatics. He is actively involved in organizing international conferences, serving, among others, as General Chair for the 2022 International Conference of Applied Intelligence and Informatics (AII2022), Local Arrangements Chair for the IEEE WCCI2020 (the world's largest IEEE CIS technical event in computational intelligence) and Publicity Chair for the 31st and 32nd International Conference on Neural Information Processing (ICONIP2024 and ICONIP2025). He received the "Hojjat Adeli Award for Outstanding Contributions in Neural Systems" (2022), awarded annually by "World Scientific Publishing" to the most innovative scientific research published in the previous year. Dr. Ieracitano is Review Editor for Frontiers in Artificial Intelligence and Associate/Academic Editor for IEEE Journal of Translational Engineering in Health and Medicine, Cognitive Computation, BMC Biotechnology and PLOS ONE. He's included since 2022 in the Top 2% researchers according to the University of Stanford/Elsevier database

Francesco Carlo Morabito is a Full Professor (2001) of Electrical Engineering and Neural Engineering with the University "Mediterranea" of Reggio Calabria, Italy. He served there as Dean (Faculty of Engineering, 2001-2008), as President of the Courses in Electronic Engineering (1998) and Industrial Engineering (2015), as Vice-Rector for Internationalisation (2013-2022), and as Deputy Rector (2017-2018). He's the founding Director of NeuroLab and AI_Lab, at DICEAM, UNIRC. He has authored or co-authored over 400 papers in international journals/conference proceedings in various fields of engineering (machine/deep learning, biomedical signal processing, radar data processing, nuclear fusion, nondestructive evaluation, artificial and computational intelligence). He has co-authored >20 intl. books (mostly focused on neural networks and machine learning) and held five national/international patents. Prof. Morabito is a Foreign Member for the Royal Academy of Doctors, Spain (2004-) and a member of the Institute of Spain, Barcelona Economic Network (2017-). Senior Member of IEEE (2000), Life SM (2024) and of INNS (2006). Governor of the International Neural Network Society (INNS), 2022-2024, and earlier for 12 years (2000-2012). President-Elect of INNS (2024), President (2025-2026). He served as President of the Italian Network Network Society (SIREN), 2008- 2014, and is co-chair of the Italian Conference on Neural Networks (WIRN). Editorial Board member for Neural Networks, International Journal of Neural Systems, and EiC for Artificial Intelligence in Neurology. He is included since 2021 in the Top 2% researchers according to the University of Stanford/Elsevier database

Amir Hussain e is Professor and founding Director of the Centre of AI and Data Science at Edinburgh Napier University, UK. His research interests are cross-disciplinary and industry-led, aimed at developing cognitive data science and AI technologies to engineer smart healthcare and industrial systems of tomorrow. He has (co)authored three international patents and over 600 publications, including around 300 journal papers, and over 20 Books/monographs. He has led major national, EU and internationally funded projects, and supervised over 35 Ph.D. students to-date. He is founding Chief Editor of Cognitive Computation journal (Springer Nature), and serves on the editorial board a various other leading journals, including Elsevier's Information Fusion and IEEE Transactions on: AI; Emerging Topics in Computational Intelligence: and Systems, Man and Cybernetics (Systems). Amongst other distinguished roles, he is an elected Executive Committee member of the UK Computing Research Committee (national expert panel of the IET and the BCS for UK computing research), General Chair of IEEE WCCI 2020 (the world's largest IEEE CIS technical event in computational intelligence, comprising IJCNN, IEEE CEC and FUZZ-IEEE), and the IEEE UK and Ireland Chapter Chair of the IEEE Industry Applications Society.

Muhammad Suffian received the M.S. degree from Mohammad Ali Jinnah University (MAJU), Karachi, Pakistan, in 2018, and Ph.D. degree from University of Urbino (UNIURB), Italy, in 2024. From 2018 to 2020, he worked as a Lecturer at MAJU and NU-FAST University Chiniot-Faisalabad, Pakistan. He was a Visiting Ph.D. Fellow at the University of Santiago de Compostela (USC), Spain. From December 2023 to August 2024, he was a Research Fellow at CTE Square Pesaro Project. He is currently Postdoctoral Research Fellow at DIIES department at the University Mediterranea of Reggio Calabria (UNIRC), Italy. His research interests include artificial intelligence, explainable artificial intelligence, computational intelligence, and biomedical data processing.

Nadia Mammone is an Associate Professor at the University Mediterranea of Reggio Calabria (Italy). She received the Laurea Degree (M.S. equivalent) in Electronic Engineering from the Mediterranea University of Reggio Calabria and the Ph.D. in ''Informatics, Biomedical and Telecommunications Engineering'' from the same University, with a dissertation that was awarded the Caianiello Prize from the Italian Neural Networks Society (SIREN). From 2014 to 2018 she was Principal Investigator at IRCCS Centro Neurolesi Bonino-Pulejo in Messina (Italy) of a research project on advanced EEG processing, funded by the Italian Ministry of Health. Formerly, she was a Post-Doc Fellow in Biomedical and Electrical Engineering at the DICEAM Department of the Mediterranea University of Reggio Calabria. She was a visiting Ph.D. Fellow at the Computational NeuroEngineering Laboratory (CNEL, University of Florida, USA) in 2005 and 2008 and a visiting Post-Doc Fellow at the Communication and Signal Processing Research Group (Department of Electrical and Electronic Engineering, Imperial College, London, UK) in 2015. In 2022 she received the ''Hojjat Adeli Award for Outstanding Contributions in Neural Systems'', awarded annually by ''World Scientific Publishing'' to the most innovative scientific research published in the previous year. She is Associate editor of Frontiers in Neuroscience (Artificial Intelligence in Neurology and Neural Technology). Her research interests include deep learning, brain computer interfaces, neural and adaptive systems, information and complex network theory. She is included since 2021 in the Top 2% researchers according to the University of Stanford/Elsevier database.