scientific reports

OPEN

E pluribus unum interpretable convolutional neural networks

George Dimas, Eirini Cholopoulou & Dimitris K. Iakovidis✉

The adoption of convolutional neural network (CNN) models in high-stake domains is hindered by their inability to meet society's demand for transparency in decision-making. So far, a growing number of methodologies have emerged for developing CNN models that are interpretable by design. However, such models are not capable of providing interpretations in accordance with human perception, while maintaining competent performance. In this paper, we tackle these challenges with a novel, general framework for instantiating inherently interpretable CNN models, named E pluribus unum interpretable CNN (EPU-CNN). An EPU-CNN model consists of CNN sub-networks, each of which receives a different representation of an input image expressing a perceptual feature, such as color or texture. The output of an EPU-CNN model consists of the classification prediction and its interpretation, in terms of relative contributions of perceptual features in different regions of the input image. EPU-CNN models have been extensively evaluated on various publicly available datasets, as well as a contributed benchmark dataset. Medical datasets are used to demonstrate the applicability of EPU-CNN for risk-sensitive decisions in medicine. The experimental results indicate that EPU-CNN models can achieve a comparable or better classification performance than other CNN architectures while providing humanly perceivable interpretations.

Recently the commercial applicability of Machine Learning (ML) algorithms has been regulated through legislation acts that aim at making the world 'fit for the digital age' with requirements, safeguards, and restrictions regarding ML and automatic decision-making in general[1]. A crucial aspect regarding the compatibility of ML models concerning these regulations is interpretability. But how is the interpretability of ML models defined? According to the recent literature[2], interpretability refers to a passive characteristic of a model, indicating the degree to which a human understands the cause of its decision. Hence, the provided interpretations of the decision-making process of a model can limit its opaqueness and earn users' trust, e.g., by offering interpretations for risk-sensitive decisions in medicine. In real-world tasks, the discriminative power of ML models, as expressed by their performance measures, e.g., their predictive accuracy, is regarded as an insufficient descriptor of their decisions[3].

Various approaches have tackled interpretability from a post hoc perspective, i.e., using methods that receive as input a fitted black-box to determine the causality of its predictions[4]. Post hoc approaches include image perturbation methods applied on the network by masking, substituting features with zero or random counterfactual instances, occlusion, conditional sampling, etc. Such approaches aim at revealing impactful regions in the image that affect the classification result[5,6]. Other post hoc methodologies handle the interpretation problem by constructing simple proxy models, with similar behavior to the original model and implement the perturbation notion at a feature-level[7,8]. This approach limits the credibility of the explanations, since the proxy model only approximates the computations of the black box[9]. Another set of techniques that reduce the complexity of operations to achieve interpretability utilize the gradient that is backpropagated from the output prediction to the input layer. These methods construct saliency maps by visualizing the gradients to present areas that are considered important by the network[10]; solely relying on their explanations, however, can be misleading[11]. In general, these methods aim at interpreting the inference of a deep learning model after its development and training, which can lead to unreliable interpretations[12]. Another drawback of applying these post-hoc methodologies to general CNN models is that the saliency maps that indicate important regions for the prediction outcome are derived from a specific latent image representation of a layer of the model. These representations are a result of the model's learning process and their interpretations are not perceptual[13].

A different approach to interpretability is the development of ML models that are interpretable by design, e.g., decision trees, lists, and sets[14]. Such models are also referred to as inherently interpretable, and usually, introduce a trade-off between interpretability and accuracy. The structure of such a model is simpler; thus, its predictive performance may be inferior to that of a more complex black-box model. However, this trade-off might be

Department of Computer Science and Biomedical Informatics, School of Science, University of Thessaly, Lamia, Greece. ✉email: diakovidis@uth.gr

preferable in high-risk decision-making domains due to the importance of understanding, validating and trusting ML models[15]. CNNs with embedded feature guiding and self-attention mechanisms in their architecture, can also be regarded as inherently interpretable[16]. These mechanisms derive interpretations by visualizing saliency maps and CNN features indicating certain concepts on the input image[17]. However, such models usually do not associate the saliency maps with human-perceivable features, and do not account for the contribution of these salient regions to the result. Other methods quantify the alignment of predefined concepts with learned filters in different layers of a network or aim towards the disentanglement of features[18] however, they do not address the direct contribution of the concept representations to the prediction[19]. Also, training such models requires a considerable manual effort for additional annotations with respect to the human-understandable concepts illustrated in each image[20]. Approaches extending regular CNNs to encode object parts in deeper convolutional layers, have also been proposed; nevertheless, they usually result in performance degradation[21]. Another approach is to leverage the intelligibility and expressiveness of Generalized Additive Models (GAMs)[22], which are recognized for their interpretability[23]. The interpretation of a GAM is based on observations associating the effect of each input feature to the predicted output. A variety of applications incorporate GAMs into their methodology to leverage their expressiveness in domains such as healthcare[24]. GAMs based on Multilayer Perceptrons (MLPs)[25], were recently proposed for interpretable data classification and regression; however, these particular models are not tailored for contemporary, CNN-based, computer vision tasks.

State-of-the-art interpretable CNN models usually exploit the information deriving from saliency maps, indicating image regions on which the model focuses its attention; however, it is not apparent how these regions contribute to the predictions. A recent relevant methodology incorporates interpretable components into a CNN model to explain its predictions[26]; nevertheless, the provided interpretations are intertwined with predefined edge kernels, and the selection of the color components does not consider any aspects of human perception. In general, there is a lack of methodologies that could explain the classification of an image based on perceptual features, i.e., features such as color and texture, described in a way that can be easily perceived and interpreted by humans[27].

Several studies have investigated ensemble models that offer interpretable results. However, most of them are targeted in stacked generalization, i.e., they combine lower-level models to minimize the generalization error for multiple predictors[28]. These studies propose ensemble models incorporating CNN sub-networks with different architectures that are trained separately by receiving the same input, or CNN sub-networks with the same architecture trained on a different partition of the same data[29]. The different CNN sub-networks are usually pre-trained on large-scale datasets. Typically, the final output is derived by some kind of a meta-learner model that combines all the individual predictions of the sub-networks of the ensemble[30,31]. The current CNN ensemble approaches rely mainly on heatmaps generated by post-hoc models to be regarded as interpretable[32,33]. More importantly, these approaches focus on improving the overall predictive accuracy, rather than implementing a unified framework that considers the interactions between each component to provide an inherently interpretable outcome.

In this paper, we propose a novel framework for the construction of inherently interpretable CNN models for computer vision tasks, motivated by the need for perceptual interpretation of image classification. The proposed framework is named after the Latin expression *E pluribus unum interpretable CNN* (EPU-CNN), which means "out of many, one" interpretable CNN. A major advantage of the proposed framework is that it is generic, in the sense that it can be used to render conventional CNN models interpretable. Given a base CNN architecture, an EPU-CNN model can be constructed as an ensemble of base CNN sub-networks, by following the GAM approach. The EPU-CNN framework requires that each sub-network of the model receives a set of orthogonal (complementary) perceptual feature representations of the same input image. EPU-CNN is therefore scalable as it can accommodate an arbitrary number of parallel sub-networks corresponding to different perceptual features. The sub-networks are jointly trained and working as one, to automatically generate interpretable class predictions. An EPU-CNN model associates the perceptual features with salient regions, as computed by the different sub-networks, and it explains a classification outcome by indicating the relative contribution of each feature to this outcome using opponent semantics. Therefore, EPU-CNN is appropriate for a wide range of application contexts, since perceptual features, such as color and texture, are essential for the discrimination of image content by humans[34]. For example, in biomedicine, abnormalities can be mainly discriminated using color features in endoscopy, texture features in magnetic resonance imaging (MRI)[35], and both color and texture features in dermoscopy for assessment of malignancies[36]. In agriculture and food quality assessment, color and texture play a decisive role for the selection of healthy plants, and for the identification of best quality fruits and meat[37,38]. Overall, there is a variety of industrial applications that could benefit from an interpretable classification system based on such perceptual features[39].

To the best of our knowledge, EPU-CNN is the first framework based on GAMs for the construction of inherently interpretable CNN ensembles, regardless of the base CNN architecture used and the application domain. Unlike current ensembles, the models constructed by EPU-CNN enables interpretable classification based both on perceptual features and their spatial expression within an image; thus, it enables a more thorough and intuitive interpretation of the classification results. Notably, ensembling shallower CNN architectures can be more efficient than training a single large model[40]. EPU-CNN, however, differentiates from other CNN ensembles both in terms of mathematical formulation and with respect to the inputs that each sub-network receives, i.e., other CNN ensembles usually propagate the same image representation to its components and each subnetwork is trained seperately[41]. Furthermore, EPU-CNN constitutes a unified framework, where all components are trained simultaneously to offer inherently interpretable classification results. In this way, EPU-CNN provides a novel approach in constructing CNN ensembles with the capacity of providing perceptually interpretable predictions. Furthermore, unlike previous interpretable CNN models[21,42], the classification performance of EPU-CNN models is comparable to or higher than that of their non-interpretable counterpart, which in the case of EPU-CNN is the base CNN model. This is demonstrated with an extensive experimental evaluation on various biomedical datasets, including datasets from gastrointestinal endoscopy and dermatology, as well as a novel contributed benchmark dataset, inspired by relevant research in cognitive science[43].

Methodology

As a framework, EPU-CNN follows the GAM approach for the construction of interpretable image classification models. GAMs represent a class of models extending linear regression models by using a sum of unknown smooth functions $\sum f_i(x_i)$, $i = 1, 2, \ldots, N$. A GAM is formally expressed as follows:

$$g(\mathbb{E}[Y|\boldsymbol{x}]) = \beta + \sum_{i=1}^{N} f_i(x_i), \tag{1}$$

where $\boldsymbol{x} = (x_1, x_{2,} \ldots, x_N)^{\mathrm{T}}$, $\boldsymbol{x} \in \mathbb{R}^N$, denotes an input feature vector, $g(\cdot)$ is a link function (e.g., logit), β is a bias term and $[Y|\boldsymbol{x}]$ denotes the expected value of the response variable Y, given an input $\boldsymbol{x}$. Each $f_i(\cdot)$, represents a univariate smooth function, $f_i : \mathbb{R} \to \mathbb{R}$, mapping each $x_i \in \mathbb{R}$ to a latent representation, $f_i(x_i)$, through which, x_i participates to the result. This structure is easily interpretable because it enables the user to explore how each input variable x_i affects the predicted output.

The EPU-CNN framework considers Eq. (1) as a template to construct an interpretable ensemble of CNNs from a conventional, non-interpretable, CNN base model (Fig. 1). The sub-networks of the ensemble are arranged in parallel, and each sub-network has the same architecture with the base model. Each sub-network receives a different input, which should be a perceptual feature representation of an input image. This representation will be referred to as *Perceptual Feature Map* (PFM) of an input image, and it can be obtained by an image transformation revealing a physical property of choice that can be easily perceived and interpreted by humans over the input image space, e.g., color and texture[27]. The number of sub-networks is determined by the number of different PFMs required to render a CNN interpretable for a particular application. Considering that each sub-network of an ensemble with a parallel topology should receive inputs with complementary information[44], the PFMs should be orthogonal. Let us consider N different PFMs I_i, $i = 1, 2, \ldots, N$, of an input image I. Each I_i is provided as input to a corresponding sub-network $\boldsymbol{C}_i(\cdot; \eta_i)$, which is parametrized by η_i, and trained jointly with the rest of the sub-networks. Hence, the input of an EPU-CNN model is a tensor $\boldsymbol{I} = (I_1, I_{2,} \ldots, I_N)$ with dimensions of $N \times H \times W$, where N, H, and W denote the number, height, and width of the PFMs, respectively. Each sub-network provides a univariate output $\boldsymbol{C}_i(I_i; \eta_i)$. The output of the EPU-CNN ensemble is computed by summing up all $\boldsymbol{C}_i(I_i; \eta_i)$, $i = 1, 2, \ldots, N$. The output of each $\boldsymbol{C}_i(I_i; \eta_i)$ can be regarded as a Relative Similarity Score (RSS), quantifying the resemblance of image I to a class with respect to I_i. Considering a binary classification problem, RSS takes values within the range of $[-1, 1]$. It represents the degree of similarity of an input image to a particular class, with respect to a particular PFM I_i. An absolute RSS value closer to 1 implies a greater similarity, whereas a positive or negative sign of the RSS associates the similarity with the one class or the other. By visualizing these scores, it becomes easier for a human to understand how each I_i affects a classification result of the EPU-CNN model. Furthermore, by examining the layer activations of $\boldsymbol{C}_i(I_i; \eta_i)$, the scores can be associated with respective image regions; thus, enabling a deeper interpretation of the classification result, based on the spatial arrangement of the observed features within the input image. The details about the PFMs considered in this study, the formulation of the classification model, and its interpretable output, are described in the following paragraphs.

Opponent perceptual feature maps. In this study the generation of PFMs is motivated by the theory of human perception of color vision proposed by Hering in the 1800's, and the opponent-process theory proposed in the 1950's by Hurvich and Jameson[45]. The behavior of a cell in the retina of the human visual system is determined by a pattern of photoreceptors, which comprises a receptive field. Receptive fields have a center-surround organization, which causes the cell to exhibit spatial antagonism, e.g., a cell that is excited by a light stimulus in the center of its receptive field will be inhibited by a light stimulus in the annulus surrounding the excitatory center. There are different types of photoreceptors, with different sensitivities to light frequency and intensity, responding differently to chromatic and luminance variations. Depending on the type of the photoreceptors, receptive fields can be color-opponent or spatially-opponent without being color-opponent[46]. Studies have provided indications that the transmitted stimuli to the retina can be decomposed into independent luminance and chromatic-opponent sources of information, and that the chromatic and luminance information of an image are processed through separate pathways by the human visual system[47,48]. Also, computer vision experiments have indicated the encoding of the chromatic and luminance components separately, as a more effective approach for image recognition[48,49].

Motivated by these studies, the proposed framework considers an opponent representation of the input images, focusing on both color and texture, which are two decisive properties for image understanding[27]. Also, color and texture provide cues enabling inferences about the shapes of objects and surfaces present in the image. Opponent color spaces have been proposed to cope with drawbacks of the RGB color space, such as the high correlation between the R, G and B color components, and its incompatibility with human perception. Representative examples include, Ohta's color space, which is obtained as a linear transformation of RGB, and it has been proposed in the context of color image segmentation, and CIE-*Lab*, which is obtained as a non-linear transformation of RGB, proposed as a device independent, perceptually uniform color space (i.e., a color space where a given numerical change corresponds to similar perceived change in color)[50]. Considering the effectiveness of CIE-*Lab* in numerous applications in computer vision, especially in biomedicine[51], in this study CIE-*Lab* is considered as a basis for the derivation of three PFMs, corresponding to its components. All the components of CIE-*Lab* are approximately orthogonal. Components *a* and *b* encode two antagonistic colors that cannot be perceived together simultaneously, e.g., there is no "reddish-green" or "bluish-yellow" color. Specifically,

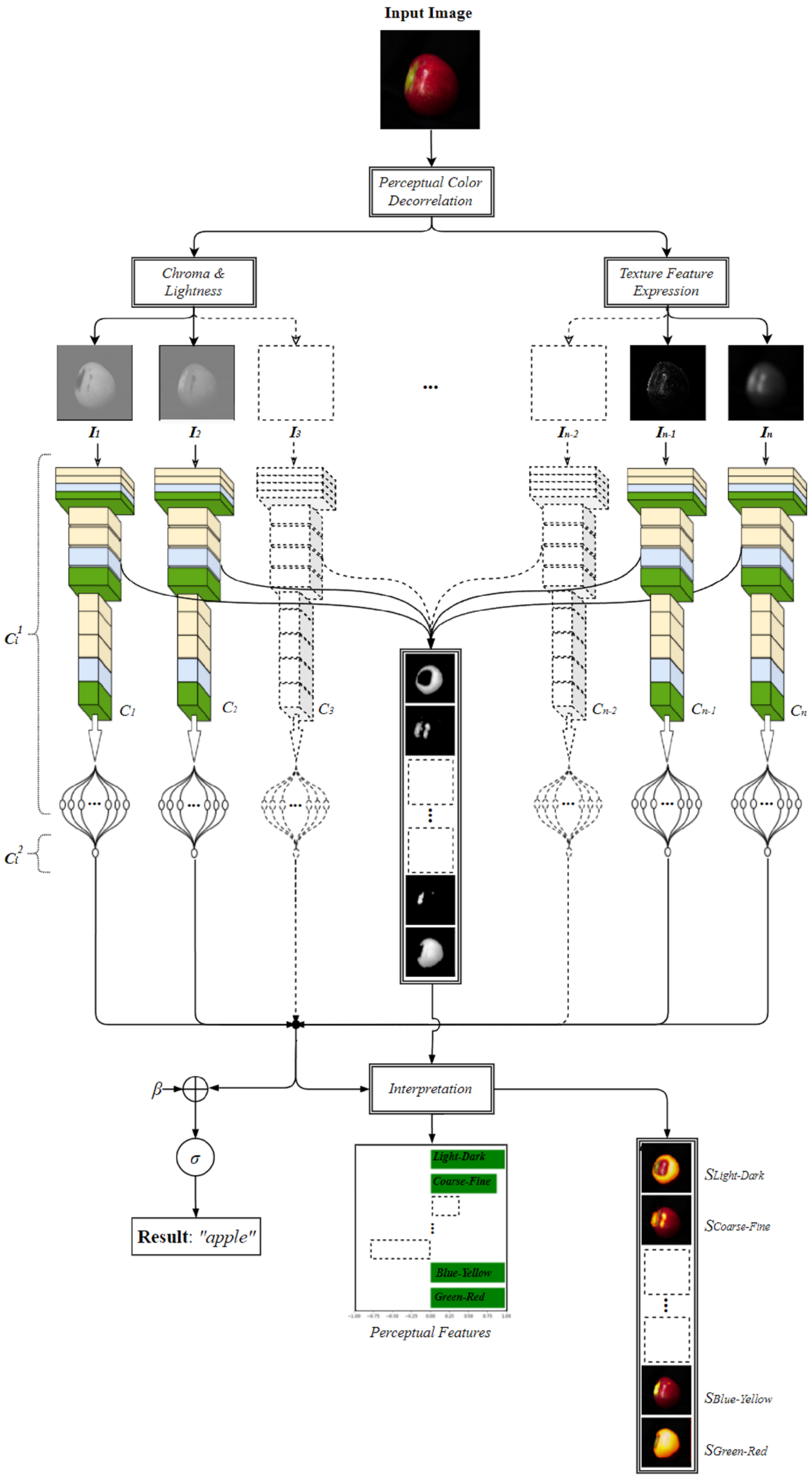


Figure 1. Outline of the EPU-CNN framework.

component a, expresses the antagonism between *green-red* hues (redness is expressed for $a > 0$, and greenness is expressed for $a < 0$), and component b expresses the antagonism between *blue-yellow* hues (yellowness is expressed for $b > 0$, and blueness is expressed for $b < 0$). The L component represents the perceptual lightness, which expresses an antagonism in luminance, between light and dark. This component, which is practically a greyscale representation of the RGB image, is usually characterized by the highest variance, as it concentrates rich information about the texture of the image contents[48].

In the field of computer vision, several studies have been based on spatial frequency representations of images, to effectively model texture for machine perception[52]. Aiming to the interpretation of the classification outcomes based on perceptual texture characteristics, the L component is further analyzed with respect to its spatial frequency. The human eye has a capacity to focus on the right range of spatial frequencies to capture the relevant details of images; thus, visual perception treats images at different levels of resolution. At lower resolutions, these details correspond to larger physical structures in a scene, whereas at higher resolutions the details correspond to smaller structures. The concept of multiresolution image representation can be modeled by the 2D Discrete Wavelet Transform (DWT)[53]. This representation is computed by decomposing the original image using a wavelet orthonormal basis. The computation of the 2D DWT is performed efficiently using the *à trous* algorithm, which is based on convolutions of the image with a pair for low and high-pass filters, called Quadrature Mirror Filters (QMFs), and dyadic down-sampling. A multilevel 2D DWT, focusing on different bands of non-overlapping spatial frequencies, can be performed by successive application of the 1-level 2D DWT on the filtered image with the lowest frequencies. The lowest frequency image of the last level represents a smooth *approximation* of the input image, where the different structures, e.g., objects, parts of objects and background, can be easier segregated upon their intensity. Based on this observation, in the context of EPU-CNN, the approximation image of the third level of the 2D DWT was selected as a PFM representing the *light–dark* antagonism with sufficiently less noise than the original L component. The higher frequency bands can be used as PFMs representing image texture at a higher detail. The selection of frequency bands depends on the application context and the pursued interpretation detail. In this study, the highest frequency band of the first level of the 2D-DWT was selected to represent the antagonistic concept of *coarse–fine* texture by exploiting the edges of the image becoming clearer at that level. The concept of coarse–fine texture is associated with the density of image edges per unit area[54], because finer textures tend to have a higher density of edges per unit area than coarser textures. Such edge-based representations are perceptually meaningful for the discrimination of different objects in a scene[55]. Considering that after each level of the 2D DWT the resolution of the image of the previous level is halved, and that the architecture of the base CNN models depends on the dimensions of the input image, the filtered images obtained after the application of the 2D DWT are up-sampled to match the size of the input image I. Thus, in this study the input tensor of an EPU-CNN model is formed as $\boldsymbol{I} = (I_1, I_2, I_3, I_4)$, where I_1 and I_2 are the PFMs corresponding to the *light–dark* and *coarse–fine* concepts respectively, $I_3 = b$ corresponds to *blue-yellow*, and $I_4 = a$ corresponds to the *green–red* concept. An example illustrating the opponent PFMs used in this study, is provided in Fig. 2.

Classification model. Given an input tensor $\boldsymbol{I}$ composed of N input PFMs, an EPU-CNN model performs feature extraction and classification. An EPU-CNN model is constructed from N CNN sub-networks $\boldsymbol{C}_i$, with each sub-network receiving a PFM $I_{i,}$ $i = 1, 2, \ldots, N$, as input. Each $\boldsymbol{C}_i$ can be regarded as a function $\boldsymbol{C}_i(\cdot; \eta_i)$, $\boldsymbol{C}_i : X^{H\times W} \rightarrow Z$, where $X^{H\times W}$ and Z are the input and univariate output space of each $\boldsymbol{C}_i$, respectively. A sub-network $\boldsymbol{C}_i$ consists of two parts: (a) a feature extractor $\boldsymbol{C}_i^1(\cdot; \omega_i)$ parametrized by ω_i; and (b) a univariate function $\boldsymbol{C}_i^2(\cdot; \theta_i)$ parametrized by θ_i. Thus Eq. (1) becomes:

$$g\left(\mathbb{E}[Y|\boldsymbol{I}; \theta_{\{N\}}, \omega_{\{N\}}]\right) = \beta + \sum_{i=1}^{N} \boldsymbol{C}_i^2\left(\boldsymbol{C}_i^1(I_i; \omega_i); \theta_i\right), \tag{2}$$

where $\boldsymbol{C}_i^1$ represents a feature extraction model composed of a CNN followed by a Fully Connected Neural Network (FC-NN), that utilizes activation functions, which are not conditioned to be smooth, and $\boldsymbol{C}_i^2$ represents a single FC-NN layer utilizing a smooth activation function that provides the final univariate output of a CNN sub-network, $\omega_{\{N\}} = \{\omega_1, \omega_2, \ldots \omega_N\}$ and $\theta_{\{N\}} = \{\theta_1, \theta_2, \ldots \theta_N\}$ are the parameters of $\boldsymbol{C}_i^1, \boldsymbol{C}_i^2$, respectively. Equation (2) encapsulates the properties and definition of GAMs while extending its capacity to exploit CNN models for computer vision tasks. The feature extractor $\boldsymbol{C}_i^1$, can be implemented by a conventional CNN architecture, whereas the number of output neurons and the activation function of $\boldsymbol{C}_i^2$ should be considered so to appropriately represent

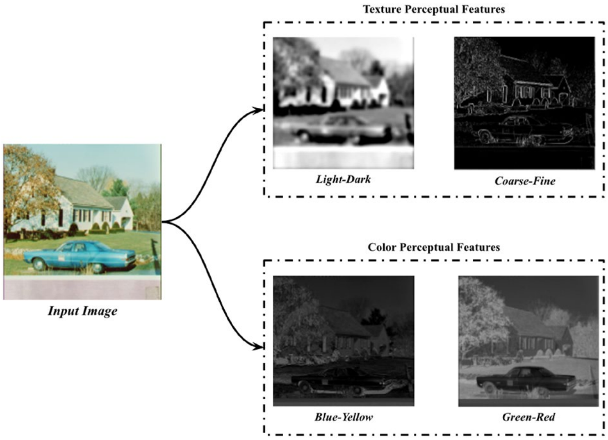


Figure 2. Illustration of the opponent perceptual features utilized by EPU-CNN.

the classification outcome, in a binary or multiclass setting. In the context of binary classification, which is considered in this study, $\boldsymbol{C}_i^2$ is formulated with a single output neuron and the hyperbolic tangent (*tanh*) activation function, resulting in sub-network responses within the range of [− 1, 1]. Ultimately, this allows to intuitively express the contribution of each feature to the final prediction, as positive or negative contribution with respect to a class label. Since EPU-CNN is applied in the context of binary classification, we chose the final output of an EPU-CNN model to be within the interval of [0, 1]. However, the formulation of EPU-CNN presented in Eq. (2) indicates that the final output can fall out of the range of [0, 1], i.e., given a N number of $\boldsymbol{C}_i$ and a bias term β the right-hand part of Eq. (2) provides values that fall within the range of $[\beta - N, \beta + N]$.

Accordingly, the $logit(\cdot)$ function, defined as:

$$logit(x) = -log\left(\frac{1}{x} - 1\right), \tag{3}$$

can be used as a suitable link function, $g(\cdot)$. Considering that the inverse of *logit* is the *log-sigmoid* function σ, Eq. (2) can be rewritten as:

$$\mathbb{E}[Y|\boldsymbol{I}; \theta_{\{N\}}, \omega_{\{N\}}] = logit^{-1}\left(\beta + \sum_{i=1}^{N} \boldsymbol{C}_i^2(\boldsymbol{C}_i^1(I_i; \omega_i); \theta_i)\right), \tag{4}$$

or

$$EPU_{CNN}\left(\boldsymbol{I}; \eta_{\{N\}}\right) = \sigma\left(\beta + \sum_{i=1}^{N} \boldsymbol{C}_i(I_i; \eta_i)\right), \tag{5}$$

where $\eta_{\{N\}} = \{\eta_1, \eta_2, \ldots, \eta_N\}$. By utilizing the *log-sigmoid* function we bound the output of EPU-CNN within the desirable range of [0, 1] suitable for binary classification applications. Equation (5) is a formal representation of an EPU-CNN model as illustrated in Fig. 1. To train an EPU-CNN model in the context of binary classification, the Binary Cross Entropy (BCE) is chosen as a loss function to be minimized:

$$EPU_{CNN} : \begin{cases} argmin_{\eta_1}\left(-\frac{1}{k}\sum_{j=1}^{k} y_j \log\left(EPU_{CNN}\left(\boldsymbol{I}_j; \eta_{\{N\}}\right)\right) + \left(1 - y_j\right)\log\left(1 - EPU_{CNN}\left(\boldsymbol{I}_j; \eta_{\{N\}}\right)\right)\right) \\ \cdot \\ \cdot \\ \cdot \\ argmin_{\eta_N}\left(-\frac{1}{k}\sum_{j=1}^{k} y_j \log\left(EPU_{CNN}\left(\boldsymbol{I}_j; \eta_{\{N\}}\right)\right) + \left(1 - y_j\right)\log\left(1 - EPU_{CNN}\left(\boldsymbol{I}_j; \eta_{\{N\}}\right)\right)\right) \end{cases}, \tag{6}$$

where $j = 1, 2, 3\ldots, k$, $EPU_{CNN}(\boldsymbol{I}_j; \eta_{\{N\}})$ is the class probability of $\boldsymbol{I}_j$ and y_j is the ground truth label of $\boldsymbol{I}_j$. As it can be observed from Eq. (6), the total error of the EPU-CNN, deriving from the responses of the CNN ensemble consensus, is used to update the parameters of each $\boldsymbol{C}_i(\cdot; \eta_i) = \boldsymbol{C}_i^2(\boldsymbol{C}_i^1(\cdot; \omega_i); \theta_i)$ of the parallel sub-network ensemble topology, simultaneously. It is worth noting that an EPU-CNN model can also be adapted for multiclass datasets, e.g., using $n > 1$ output neurons instead of one, in the case of $n > 1$ classes. Then, the network's output can be interpreted by considering the contribution of the multiclass classification outcome of each CNN sub-network to the final classification result (see "Qualitative comparison with state-of-the-art interpretable methods" section).

Interpretable output. Given an input image, EPU-CNN provides three outputs, as illustrated in Fig. 1, namely, (a) the predicted class $EPU_{CNN}(\boldsymbol{I}; \eta_{\{N\}})$; (b) a set of RSSs $\boldsymbol{C}_i(I_i; \eta_i)$, $i = 1, 2, \ldots, N$, explaining why the image is classified in that class; and (c) a set of Perceptual Relevance Maps (*PRMs*) S_i explaining which image regions are responsible for each RSS. Figure 3 illustrates the provided outputs of the model for two images that belong to different classes. The classification result is indicated as a textual label characterizing the input image, and the RSSs are visualized through bar-charts. Each bar-chart consists of horizontal red or green colored bars, indicating the magnitude of resemblance that each I_i is estimated to have for the banana and apple class, respectively. Additionally, the model provides with respect to each I_i, areas (PRMs) highlighting their resemblance to the predicted class. The color scaling from orange to yellow regions of the maps indicates the ascending intensity of activation.

Image-specific visualizations of RSSs enable the interpretation of the classification process of unlabeled input images. This is the most important aspect of an EPU-CNN model. For example, the image of Fig. 3a, is classified as a banana, because all PFMs, i.e., *light–dark*, *coarse–fine*, *blue–yellow* and *green–red*, as indicated by the respective RSSs, guide the prediction towards the banana class, which corresponds to negative $\boldsymbol{C}_i(I_i; \eta_i)$ responses (red). Accordingly, the image of Fig. 3b, is classified as an apple, because all PFMs guide the prediction towards the apple class, i.e., positive $\boldsymbol{C}_i(I_i; \eta_i)$ responses (green). However, it is not necessary for all the $\boldsymbol{C}_i(I_i; \eta_i)$ responses to be negative for an image to be classified as a banana, since EPU-CNN models consider the consensus of the sub-networks.

Perceptual Relevance Maps, S_i, are generated to visually inspect the relevant regions of the input image I with respect to each RSS $\boldsymbol{C}_i(I_i; \eta_i)$. Let $F_i^l = (f_{i,l}^1, f_{i,l}^2, f_{i,l}^3 \ldots, f_{i,l}^n)$ indicate a tensor of feature maps with $F_i^l \in \mathbb{R}^{n \times h \times w}$, where n, h, w denote the depth, height and width of F_i^l, and $f_{i,l}^n \in \mathbb{R}^{h \times w}$, as computed by a convolutional layer l of a $\boldsymbol{C}_i$. The selection of l is intertwined with its capacity to highlight regions that contribute to the derivation of $\boldsymbol{C}_i(I_i; \eta_i)$. The deeper the layer l that F_i^l is extracted from, the more approximate the correspondence among the

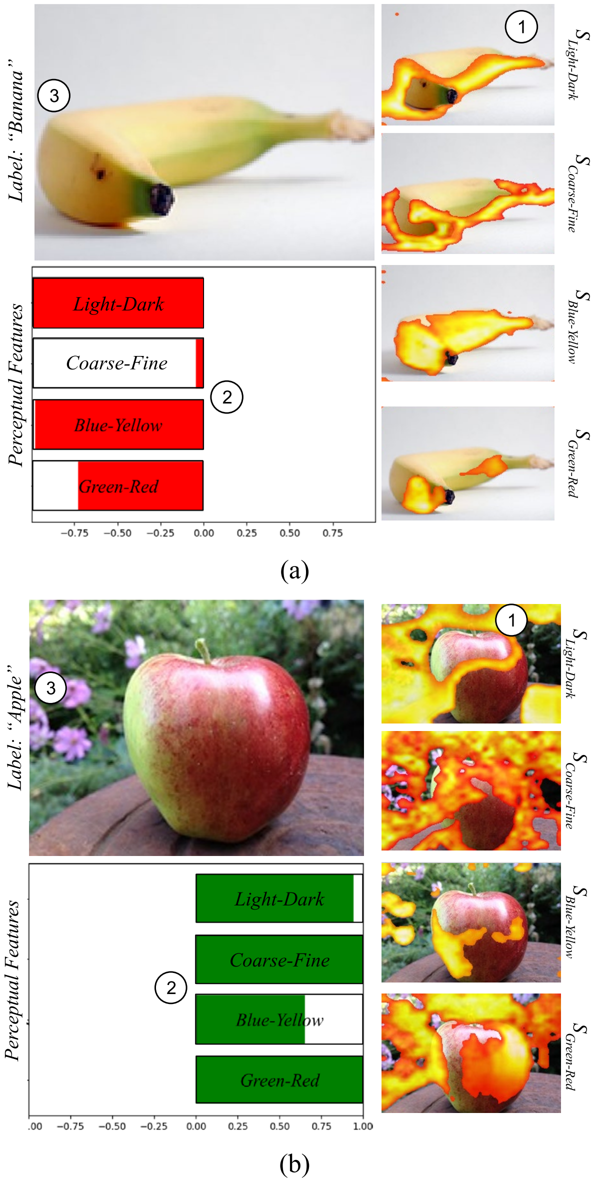


Figure 3. Example of EPU-Net output visualization using bar-charts and saliency maps. The numbering indicates the interpretation order of EPU-CNN output. The label field indicates the predicted label. (**a**) Interpretation of an image classified as a banana. (**b**) Interpretation of an image classified as an apple.

feature maps $f^n_{i,l}$ and the input image I; thus, a middle layer l of $\mathbf{C}_i$ is considered for the construction of each S_i[56] ("Ablation study" section).

To quantify the amount of information that each $f^n_{i,l}$ encodes, we compute the Shannon Entropy (SE) scores. Then, half of the most informative $f^n_{i,l}$, *i.e.*, $f^n_{i,l}$ that correspond to the highest entropy scores, are aggregated to construct the S_i. The aggregation is performed by averaging $f^n_{i,l}$ features maps which results to the initial S_i estimation. Then S_i is further refined, by applying a thresholding method that maximizes the entropic correlation between the foreground and background of S_i, for maximum information transfer[57]. The entropy-based thresholding operation is performed to exclude values associated with lower saliency and communicate to the user the most informative regions. An example of different S_i of input images I can be seen in Fig. 3. The generated S_i illustrated in Fig. 3 are overlayed on the input images (Fig. 3a,b). The highlighted regions indicate the spatial association of similarity scores $\mathbf{C}_i(I_i; \eta_i)$ with the respective input image. Moreover, the numbers in the images of Fig. 3 indicate in which order the different outputs of an EPU-CNN can be considered by the user. Initially a user can examine the regions that are highlighted by the generated PRMs of each PFM (1). Subsequently, these regions are participating to the classification outcome, towards either class, with a magnitude that is indicated by the RSSs (2). Finally, the PRMs (1) along with the RSSs (2) can assist the user to interpret the class prediction of an EPU-CNN (3).

Experiments and results

Datasets. EPU-CNN was trained and evaluated on nine different datasets. Initially, a dataset specifically created for the evaluation of the interpretability capabilities of EPU-CNN was considered. The purpose of using this dataset was to demonstrate the capabilities of EPU-CNN with clear, simple, and perceptually meaningful examples. Considering biomedicine as a critical application area for explainable and interpretable artificial intelligence (AI), four well-known biomedical benchmark datasets, consisting of endoscopic and dermoscopic images, was used for further evaluation. Furthermore, to demonstrate the generality of the proposed approach, four well-known benchmark datasets for real image classification were considered.

Interpretability dataset. For the purposes of this study, a novel dataset was constructed, named Banapple. The dataset consists of images of bananas and apples. It was created by collecting images, under the Creative Commons license, from Flickr. The images illustrate bananas and apples with variations regarding the color, placement, size, and background. The motivation for the construction of this dataset stems from studies in cognitive science, where human perception is investigated using examples with discrete properties of bananas and apples[43]. The experiments performed aim to demonstrate that EPU-CNN is capable of capturing the discriminative characteristics of bananas and apples by the perceptual features it incorporates, i.e., apples have a circular shape and usually red color, whereas bananas have a bow-like shape and usually a yellow color. In addition, samples that deviate from the average appearance of these objects can provide insights regarding the reliability of the interpretation of the model.

Endoscopic datasets. Publicly available datasets of endoscopic images were considered for the evaluation process. Namely, KID[58], Kvasir[59] and a dataset that was part of the MICCAI 2015 Endovis challenge[60]. The KID dataset consists of 2352 annotated wireless capsule endoscopy (WCE) images of abnormal findings i.e., inflammatory, vascular and polypoid lesions as well as images depicting normal tissue from the esophagus, stomach, small bowel and colon. The Kvasir dataset consists of images of the gastrointestinal (GI) tract, annotated and verified by medical experts. These include 4000 images of anatomical landmarks, i.e., Z-line, pylorus and cecum, and pathological findings of esophagitis, polyps and ulcerative colitis. The dataset also contains sets of images related to endoscopic polyp removal that were not utilized for this work. The MICCAI 2015 Endovis challenge dataset consists of 800 gastroscopic images of normal and abnormal findings, such as gastritis, ulcer, and bleeding.

Dermoscopic dataset. The evaluation process of EPU-CNN has also included the International Skin Image Collaboration Challenge 2019 (ISIC2019) dermoscopic image collection. ISIC2019 challenge provides a publicly available archive of 25,331 dermoscopic images of eight different categories of skin lesions, namely, melanoma, melanocytic nevus, carcinomas (both of basal and squamous cells), actinic and benign keratosis, dermatofibroma, and vascular lesions. These images were used to construct three different binary classification problems: (a) melanomas *vs.* melanocytic nevi (*Me. vs. Ne.*); (b) carcinomas *vs.* melanocytic nevi (*Ca. vs. Ne.*) and (c) carcinomas *vs.* melanomas.

(Ca. vs. Me.). The tasks (a) and (b) are characterized as a classification between abnormal (carcinomas, melanomas) and normal (melanocytic nevus) skin lesions whereas task (c) discriminates two abnormal categories of different incidence and survival rates, i.e., melanomas have higher mortality rates than carcinomas[61]. Task (a) comprised of 9000 images whereas task (b) and (c) 8200 and 8500 images respectively.

To demonstrate the generality of the proposed framework, EPU-CNN was further validated on well-recognized non-biomedical benchmark datasets CIFAR-10, MNIST, Fashion MNIST and iBean.

CIFAR-10. CIFAR-10 consists of 60,000 color images of natural objects that belong to 10 different classes. The dataset is split into 50,000 training and 10,000 test images and each class comprises 6000 images with a size of 32 × 32 pixels.

MNIST. MNIST is a database of handwritten digits, that consists of a training set of 60,000 images, and a test set of 10,000 images. Each image has a size of 28 × 28 pixels, representing a single digit in greyscale among 10 classes (from 0 to 9). The dataset follows a balanced distribution as each digit is represented by 6000 images for training and 1000 images for testing.

Fashion MNIST. Fashion MNIST is a database of fashion items that consists of a training set of 60,000 images and a test set of 10,000 images. Each image has a size of 28 × 28 pixels, associated with a label from 10 classes, representing different fashion items in greyscale, such as T-shirts, shoes, dresses etc. The dataset follows a balanced distribution as each fashion item is represented by 6000 images for training and 1000 images for testing.

iBean. iBean is a dataset of color images that was created for the classification of diseases in bean plants. It consists of 1295 images of bean leaves, that belong to three different classes. The classes represent 428 healthy leaves, 432 leaves with angular leaf spot disease, and 436 leaves with bean rust disease.

Classification performance assessment. For the comparison of the classification performance of EPU-CNN, we selected three well established CNN models, namely, VGG_{16}[62], $ResNet_{50}$[63] and DenseNet_{169}[64] and an inherently interpretable CNN model abbreviated as TT[65]. In this study, the VGG_{16} was used as a base for the TT model. The same training parameters, i.e., batch size, optimization algorithm and data augmentation, were applied on all networks involved in the evaluation process. In detail, the batch size was set to 64 and as an optimization algorithm the Stochastic Gradient Decent was used; the training data were augmented only with respect to their orientation. The weights of all networks were randomly initialized before training. Five different CNNs architectures were considered for the construction of EPU-CNN models. More specifically, two indicative CNN architectures, namely, $Base_{I}$ and $Base_{II}$, along with VGG$_{16}$, ResNet$_{50}$ and DenseNet$_{169}$ were incorporated as base models in the EPU-CNN framework. These models were selected to demonstrate the generality of the proposed framework, i.e., its applicability to rendering different conventional CNN architectures interpretable. Regarding the architecture of the indicative CNN architectures, $Base_{I}$, consists of 3 convolutional blocks in total, followed by an FCNN. The first two convolutional blocks are identical and include two convolutional layers followed by a max-pooling and a batch normalization layer. The convolutional layers of these blocks have a depth size of 64 and 128 respectively. The following convolutional block consists of three convolutional layers with a depth size of 256 followed by a max-pooling and a batch normalization. All the kernels of the convolutional layers had a size of 3 × 3. $Base_{II}$, follows the same architecture with $Base_{I}$ with an additional convolutional block, in the beginning of the architecture, utilizing an inception module. $Base_{I}$, $Base_{II}$, VGG$_{16}$, ResNet$_{50}$ and DenseNet$_{169}$ were used for the construction of EPU_{I}, EPU_{II}, EPU_{VGG}, EPU_{ResNet} *and* $EPU_{DenseNet}$, respectively.

The evaluation followed a tenfold cross validation procedure with the average Area Under the receiver operating Characteristic (AUC) score among all folds. The AUC was selected as an overall summary measure of binary classification performance, which unlike accuracy, is relatively robust for datasets with imbalanced class distributions[66]. The performance of all models is summarized in Table 1. The best results are in boldface typesetting and the results ranked second are underlined. It can be observed that the results obtained by the EPU-CNN models indicate an overall better or comparable classification performance to their non-interpretable counterparts, i.e., $Base_{I}$, $Base_{II}$, VGG$_{16}$, ResNet$_{50}$ and DenseNet$_{169}$. More specifically, on Banapple, Endovis, Kvasir and ISIC 2019 (*Me. vs. Ne.*). In most cases EPU_{II} provided better results when compared with the other EPU-CNNs and with the majority of the base models. However, although in some cases the EPU-CNN models did not provide a better classification accuracy than the base models, their advantage is that their output is interpretable. The interpretability of these models is assessed in the following sub-section.

Figure 4 illustrates the number of trainable parameters of each model. It can be observed that the complexity of an EPU-CNN model is analogous to that of its base model. Additionally, an EPU-CNN can provide competent results even with base models of low complexity, i.e., EPU_{I} and EPU_{II} utilize ~ 40 and ~ 19 million parameters, respectively; however, they provide higher classification performance when compared to more complex EPU-CNN models. Furthermore, the less complex EPU_{II} has comparable or even better classification performance when compared to EPU_{I} while it outperforms substantially ResNet$_{50}$ that is a more computationally demanding base model (~ 27 M parameters). On the other hand, the inherently interpretable (TT) model that has a similar complexity to EPU_{II}, i.e., ~ 20 M parameters, provides the lowest overall classification performance amongst all models.

Quantitative interpretability analysis. To quantitatively evaluate the interpretability of the proposed framework we exploited the properties of the Banapple benchmark dataset. Banapple is suitable for this purpose because our perception of the class-related objects is directly associated with the way we categorize them, based on their visual attributes regarding color and shape[43]. Therefore, the subsequent task of annotating the images of Banapple did not require any domain-specific knowledge. Images of bananas and apples have distinguishable characteristics with respect to all PFMs utilized by the EPU-CNN models, i.e., *light–dark*, *coarse–fine*, *blue-yellow* and *green–red*. Thus, in the case of a correct class prediction, ideally, all RSSs should trend towards the same direction, as indicated by the sign of an RSS, e.g., all RSSs for an apple image should be positive, whereas for a banana image should be negative. Hence, given that the EPU-CNN models in this study use four PFMs, a ground truth, $\boldsymbol{y}_{int}$, and predicted, $\widetilde{\boldsymbol{y}}_{int}$, interpretability label is expressed as follows:

	Datasets						
					ISIC 2019		
Models	***Banapple***	***KID***	***Endovis***	***Kvasir***	***Ca.vs.Ne***	***Ca.vs.Me***	***Me.vs.Ne***
EPU_{I}	0.91 ± 0.01	0.94 ± 0.02	*0.97 ± 0.01*	0.87 ± 0.02	0.96 ± 0.03	*0.92 ± 0.01*	*0.94 ± 0.02*
EPU_{II}	**0.92 ± 0.01**	0.93 ± 0.01	**0.97 ± 0.01**	**0.92 ± 0.01**	0.94 ± 0.02	0.91 ± 0.01	**0.94 ± 0.01**
EPU_{VGG}	0.90 ± 0.01	0.93 ± 0.01	0.89 ± 0.03	0.88 ± 0.01	**0.97 ± 0.04**	0.90 ± 0.01	0.89 ± 0.03
EPU_{ResNet}	0.84 ± 0.04	0.86 ± 0.06	0.84 ± 0.01	0.79 ± 0.04	0.88 ± 0.08	0.78 ± 0.06	0.86 ± 0.04
$EPU_{DenseNet}$	0.90 ± 0.03	0.93 ± 0.05	0.87 ± 0.01	0.90 ± 0.03	*0.97 ± 0.03*	0.92 ± 0.02	0.92 ± 0.03
$Base_{I}$	*0.92 ± 0.02*	*0.96 ± 0.02*	0.96 ± 0.01	*0.91 ± 0.01*	0.90 ± 0.02	**0.94 ± 0.03**	0.93 ± 0.02
$Base_{II}$	0.91 ± 0.01	**0.97 ± 0.01**	0.97 ± 0.01	0.91 ± 0.01	0.93 ± 0.01	0.92 ± 0.04	0.93 ± 0.04
VGG_{16}	0.90 ± 0.01	0.90 ± 0.04	0.93 ± 0.01	0.85 ± 0.01	0.87 ± 0.01	0.90 ± 0.03	0.92 ± 0.02
$ResNet_{50}$	0.89 ± 0.02	0.92 ± 0.03	0.88 ± 0.10	0.87 ± 0.09	0.69 ± 0.12	0.90 ± 0.02	0.92 ± 0.03
$DenseNet_{169}$	0.88 ± 0.04	0.94 ± 0.05	0.90 ± 0.12	0.88 ± 0.01	0.76 ± 0.09	0.90 ± 0.01	0.91 ± 0.03
TT_{VGG}	0.82 ± 0.04	0.91 ± 0.03	0.93 ± 0.05	0.85 ± 0.04	0.88 ± 0.03	0.76 ± 0.01	0.81 ± 0.02

Table 1. Classification results (AUC) of EPU-CNN and CNN models. Significant values are in bold and italics.

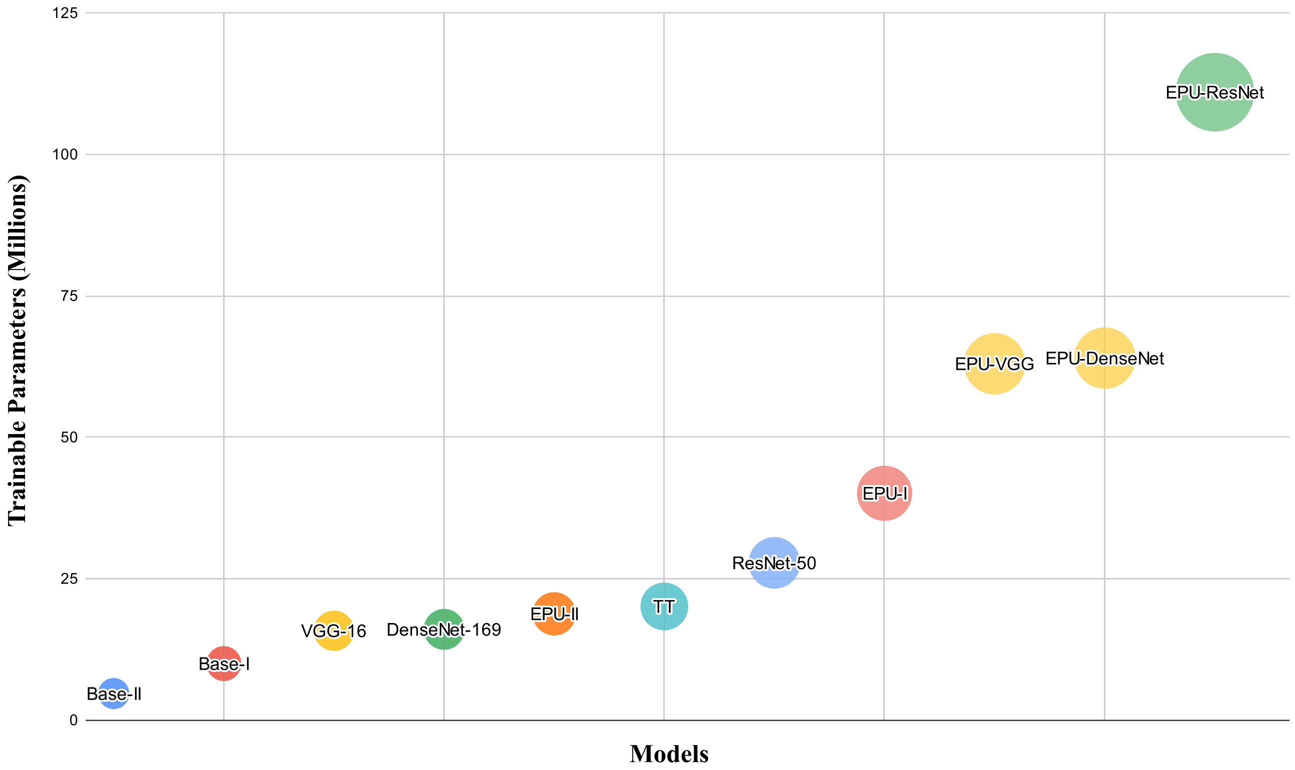


Figure 4. Visualization of the complexity of the compared models in terms of the number of trainable network parameters.

$$\boldsymbol{y}_{int}(y) = \begin{cases} (1,1,1,1), if y = 1 \\ -(1,1,1,1), if y = 0 \end{cases}, \tag{7}$$

$$\widetilde{\boldsymbol{y}}_{int}\left(EPU_{CNN}\left(\boldsymbol{I};\ \eta_{\{n\}}\right)\right) = \left(sign(\boldsymbol{C}_1(I_1;\ \eta_1)), \ldots, sign(\boldsymbol{C}_4(I_4;\ \eta_4))\right), \tag{8}$$

where y is the ground truth class label of an image I, 1 and 0 denotes the apple and banana class respectively whereas $sign(\boldsymbol{C}_i(I_i;\ \eta_i))$ returns the sign of an RSS. Given a set of ground truth and predicted interpretability label pairs, the interpretability accuracy a_{int}, is calculated as the average Jaccard Index[67], $J(\cdot)$, among them, as follows:

$$a_{int} = \frac{1}{k}\sum_{j=1}^{k} J\left(\boldsymbol{y}_{int}(y_j), \widetilde{\boldsymbol{y}}_{int}\left(EPU_{CNN}\left(\boldsymbol{I}_j;\ \eta_{\{n\}}\right)\right)\right). \tag{9}$$

The results reported in Table 2 show that EPU_I and EPU_{II} achieved the highest a_{int} with a score of 72.40 ± 1.51% and 72.62 ± 1.63% respectively. This means that the capacity of both EPU_I and EPU_{II} models is comparable with respect to their capacity to interpret the classification of bananas and apples. Since EPU_{II} achieves a better overall classification performance, a_{int} score and it is more computationally efficient, it has been chosen for the qualitative investigation of interpretability that is presented in the following sections.

Ablation study.

An ablation study was conducted to assess the impact of each component predictor (sub-network) of EPU-CNN for all possible combinations of the PFMs that have been considered in this study, i.e., *light–dark* (LD), *coarse–fine* (CF), *blue-yellow* (BY) and *green–red* (GR). The results are reported in Table 3 in terms of AUC scores. It can be noticed that the best results for the Banapple were obtained using all four PFMs.

Metric	EPU-CNN models				
	EPU_I	EPU_{II}	EPU_{VGG}	EPU_{ResNet}	$EPU_{DenseNet}$
$a_{int\ (\%)}$	72.40 ± 1.51	72.62 ± 1.63	66.64 ± 2.21	62.90 ± 4.13	64.62 ± 2.24

Table 2. Interpretability accuracy results of EPU-CNN models.

	Datasets		
Components	*Banapple*	*Endoscopic*	*Dermoscopic*
GR	0.77 ± 0.03	0.89 ± 0.02	0.87 ± 0.01
BY	0.84 ± 0.02	0.80 ± 0.03	0.86 ± 0.01
CF	0.86 ± 0.03	0.78 ± 0.03	0.89 ± 0.01
LD	0.76 ± 0.02	0.77 ± 0.04	0.86 ± 0.02
GR-BY	0.89 ± 0.01	**0.96 ± 0.01**	0.90 ± 0.01
GR-CF	0.84 ± 0.02	0.95 ± 0.02	0.91 ± 0.03
GR-LD	0.82 ± 0.01	0.91 ± 0.03	0.90 ± 0.02
BY-CF	0.91 ± 0.02	0.86 ± 0.03	0.92 ± 0.02
BY-LD	0.87 ± 0.02	0.84 ± 0.02	0.90 ± 0.02
CF-LD	0.88 ± 0.01	0.82 ± 0.03	0.91 ± 0.03
BY-CF-LD	0.91 ± 0.02	0.85 ± 0.03	**0.94 ± 0.01**
GR-CF-LD	0.89 ± 0.01	0.92 ± 0.02	0.92 ± 0.02
GR-BY-LD	0.88 ± 0.02	0.93 ± 0.01	0.91 ± 0.01
GR-BY-CF	0.91 ± 0.01	0.95 ± 0.02	0.92 ± 0.02
GR-BY-CF-LD	**0.92 ± 0.01**	0.92 ± 0.01	0.93 ± 0.01

Table 3. AUC scores of the ablation study on all PFM combinations. Significant values are in bold.

Considering the Dermoscopic datasets, the best results were obtained using PFMs of *light–dark*, *coarse–fine* and *blue-yellow*, and for the endoscopic dataset best results were obtained using only the color PFM representations.

An additional ablation study was performed to determine the impact of layer selection to the construction of PRMs using EPU_{II} as the best performing model. The feature maps estimated by 3 different layers have been chosen for the construction of the respective PRMs. Each of these layers corresponded to the last layer of each convolutional block of EPU_{II}. Figure 5 illustrates indicative PRMs constructed using feature maps estimated by different convolutional layers on predictions from the Banapple, Endoscopic, and Dermoscopic datasets. As it can be observed, the regions identified as meaningful regarding each PFM are approximately consistent with each other regardless of the degree of abstraction that each set of feature maps encodes. However, the feature maps estimated by the intermediate 5th layer provide less noisy PRMs that highlight with more precision the areas on the input image that are estimated to be meaningful with respect to each PFM.

Qualitative interpretability analysis. The qualitative analysis of EPU-CNN was investigated by considering both PRMs, global and local bar-charts generated by the EPU_{II} model, for each dataset. Given a validation set of images with a priori known class memberships, global bar-charts are constructed by averaging the RSSs per class, as provided by each sub-network of EPU_{II}. Global bar-charts enhance the transparency of the model and reveal the overall contribution of PFMs regarding the data discrimination process. In a global bar-chart, PFMs of low or high significance can be identified by their dataset-wide score, which can lead to the selection of a subset of the most informative PFMs, i.e., by pruning or replacing the sub-networks the PFMs of low significance. The respective results obtained per dataset are provided in the next paragraphs.

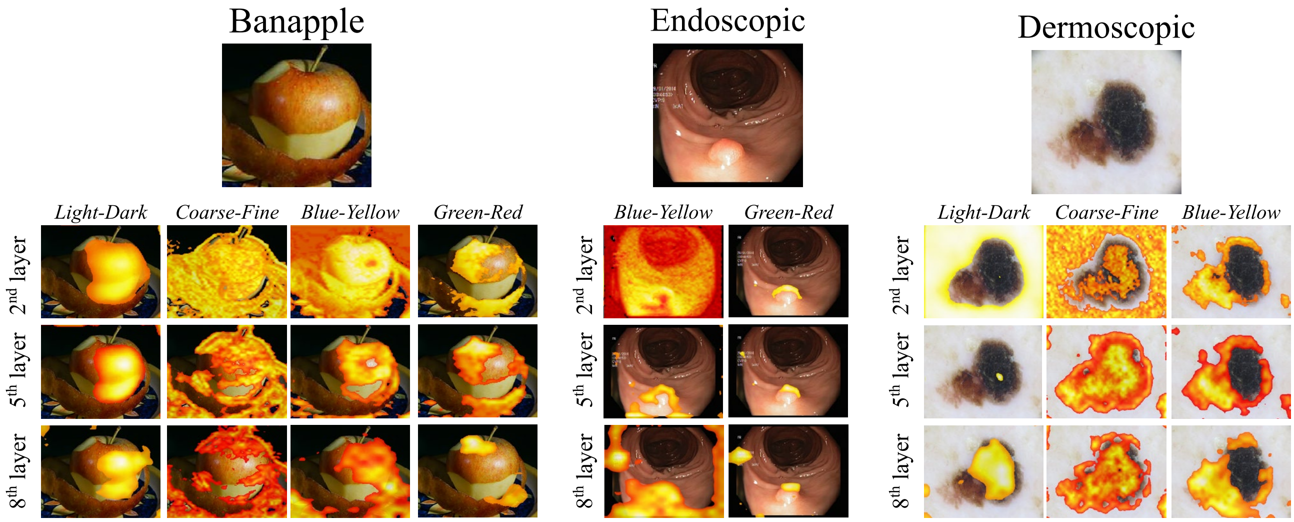
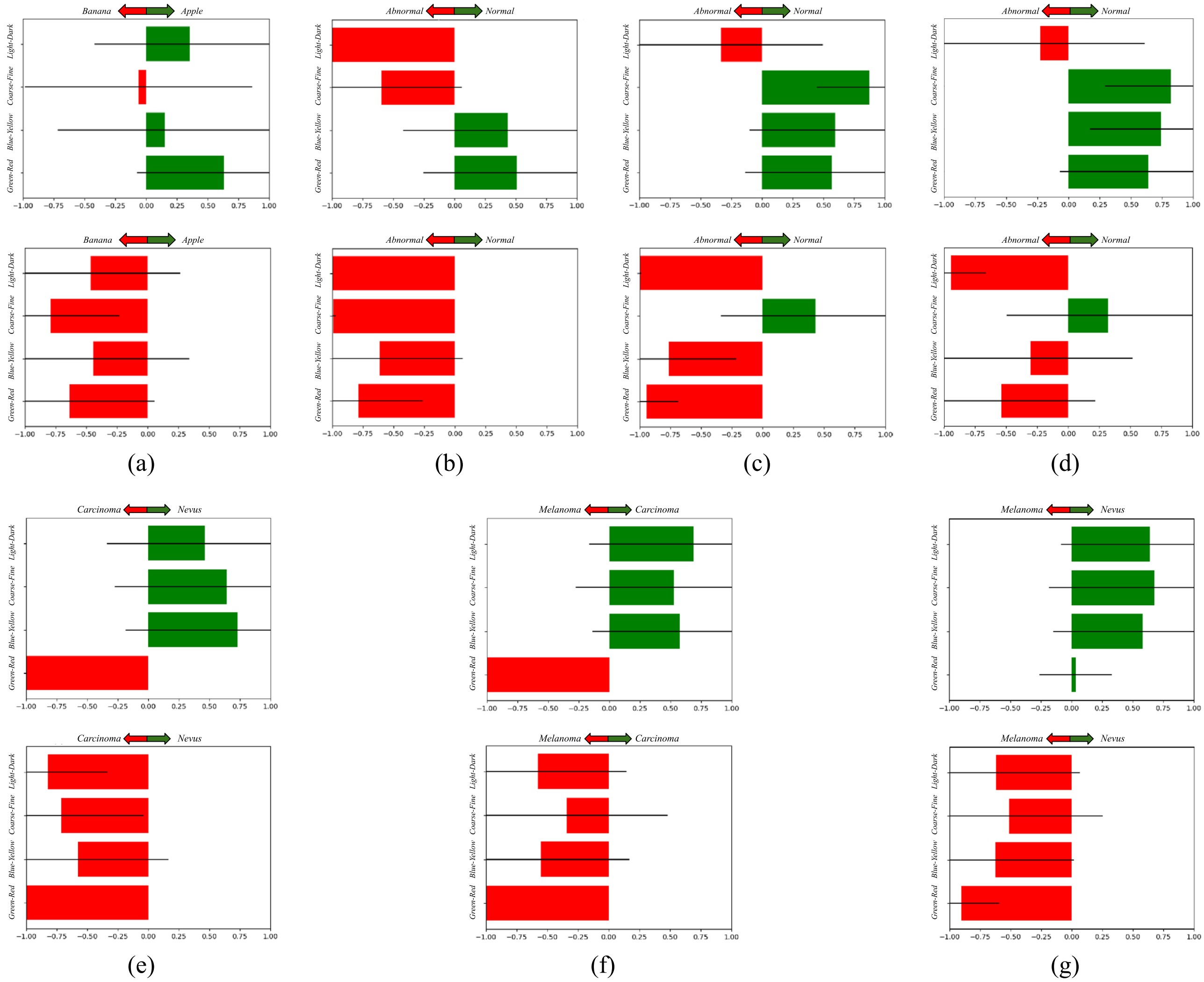


Figure 5. Example of PRMs generated by features maps extracted from different layers of EPU_{II}.

Banapple. The global bar-charts illustrated in Fig. 6a indicate that all the perceptual features contribute to the classification of the images. This result is in accordance with our perceptual understanding[43], since apples and bananas are discriminated with respect to all PFMs considered in this study. Figure 7 illustrates examples of local bar-charts along with the respective PRMs of classified images. Specifically, the images presented in the first column of Fig. 7, were correctly classified by EPU_{II}, and this is reflected by the visualization of the RSSs. The PRMs of each sub-network indicate the regions of the input image which resemble the class that each RSS suggests. For instance, in Fig. 7d the highlighted areas of the PRMs corresponding to *green–red* and *blue-yellow* are overlayed precisely on the class-related object, i.e., the bananas. Interestingly, the difference between the *light–dark* and *coarse–fine* RSSs can be justified by the obscurity of the highlighted regions of $S_{light\text{-}dark}$ and $S_{coarse\text{-}fine}$, i.e., both PRMs highlight the table.

The second column of Fig. 7 illustrates wrongly classified images. Notably, each of these images have resemblances to the opposite class with respect to color and shape. For example, in Fig. 7b the perceptual features of *light–dark*, *coarse-fine*and *blue-yellow*, wrongfully guide the prediction towards the banana class (red). This can be justified since the image contains objects that share characteristics that resemble the banana class, i.e., the shape and color of the hands holding the apple. The RSS of *green–red* however, trends towards the apple class (green) with high magnitude, whereas the respective $S_{green\text{-}red}$, highlights the apple. Accordingly, the PRMs of *light–dark* and *coarse–fine* focus on the hands explaining the trend of the respective RSSs towards the banana class. Nevertheless, even though $S_{blue\text{-}yellow}$ focuses on the apple, the respective RSSs indicate that the image belongs to the banana class. In Fig. 7e the *light–dark* and *blue-yellow* RSSs trend towards the apple class (green). The direction of these RSSs towards the incorrect class can be justified considering that the color and orientation of the bananas are not representative of their class. Accordingly, $S_{light\text{-}dark}$ and $S_{blue\text{-}yellow}$ focus only partially on the banana. Similarly, the shape and color from the inside of the apple in Fig. 7h are unusual for an apple. Hence, the



(a) (b) (c) (d)

(e) (f) (g)

Figure 6. Example of dataset-wide interpretations provided by EPU-CNN on all datasets. Green (positive response) and red (negative response) bars indicating participation the 1 and 0 class respectively, and the black lines indicate the standard deviation. (**a**) Banapple. (**b**) KID. (**c**) MICCAI Endovis 2015. (**d**) Kvasir. (**e–g**) ISIC 2019.

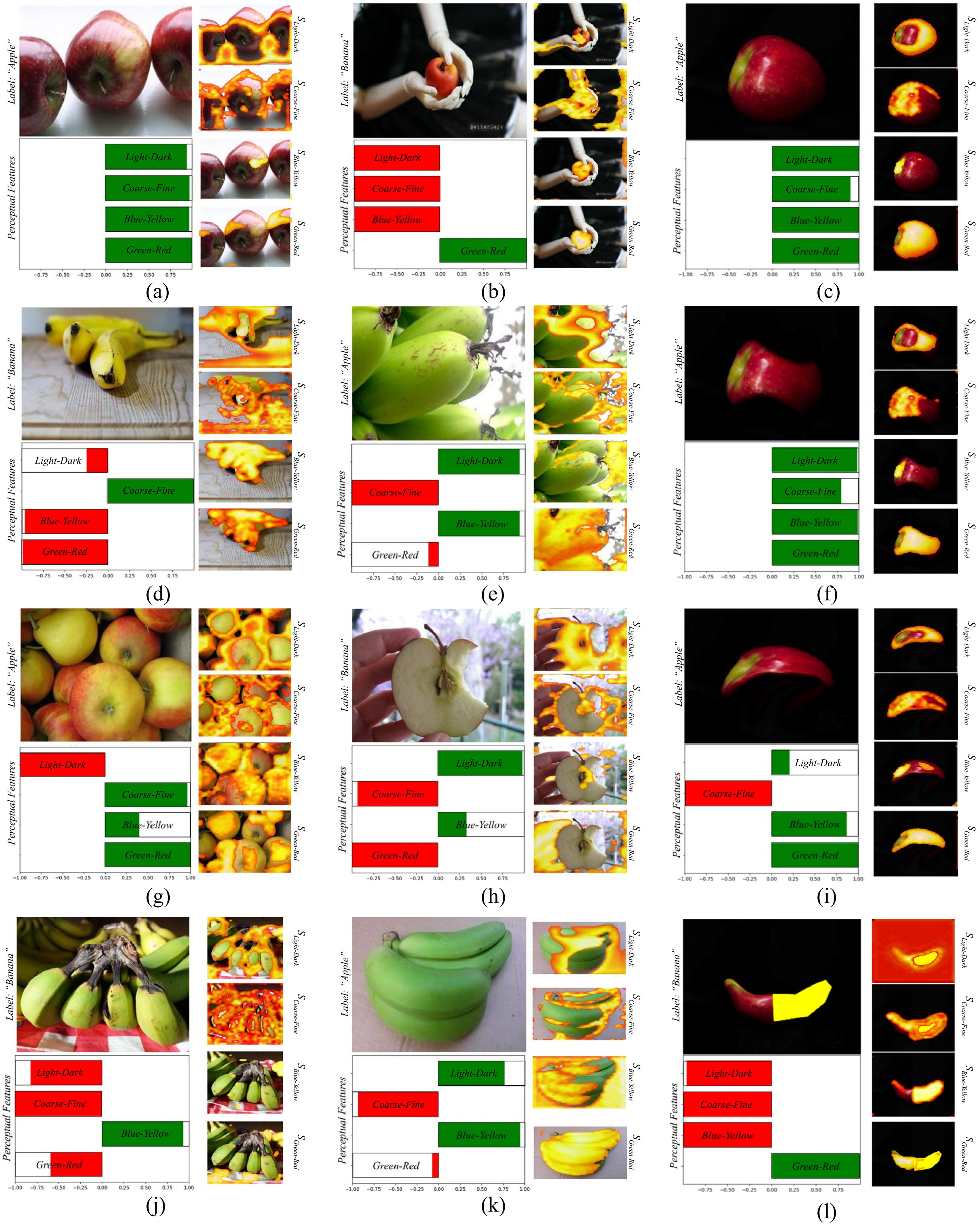


Figure 7. Example of local bar-charts produced by EPU_{II} on images from the Bananapple dataset. The label field indicates the predicted label. First column shows correctly classified images. Second column shows wrongly classfied images. Third column shows changes in the classification and its interpretation.

coarse–fine and *red-green* RSSs lean towards the opposite direction. It can be observed that the negative *red-green* and *coarse–fine* scores, have corresponding PRMs that do not focus on the class-related object, i.e., they highlight regions of the hand and the background. Also, the greenish color of the bananas in Fig. 7k, can be descriptive for both classes (as both bananas and apples can be green), which is also expressed by the disagreement between the relative scores of the color PFMs. Interestingly, the disagreement between the *light–dark* and *coarse–fine* scores can also be justified by the highlighted regions in the respective PRMs, *i.e.,* the outline of the banana object in $S_{coarse\text{-}fine}$ and the circular region, resembling an apple in $S_{light\text{-}dark}$.

To further investigate the behavior of EPU-CNN, we have chosen an image depicting an apple (Fig. 7c) which was digitally processed to obtain 3 variations: (a) to illustrate a bitten apple (Fig. 7f); an apple with a shape resembling that of a banana (Fig. 7i); and (c) an apple resembling both the shape and color of a banana while maintaining a reddish region (Fig. 7l). The interpretation changes that can be observed include the following:

(a) When the shape resembles a bitten apple the *coarse–fine* PFM is still guiding the prediction towards the apple class but with greater uncertainty (Fig. 7f), whereas the $S_{coarse\text{-}fine}$ discriminates the image based on its textural variations, i.e., the curvature of the left side of the apple.

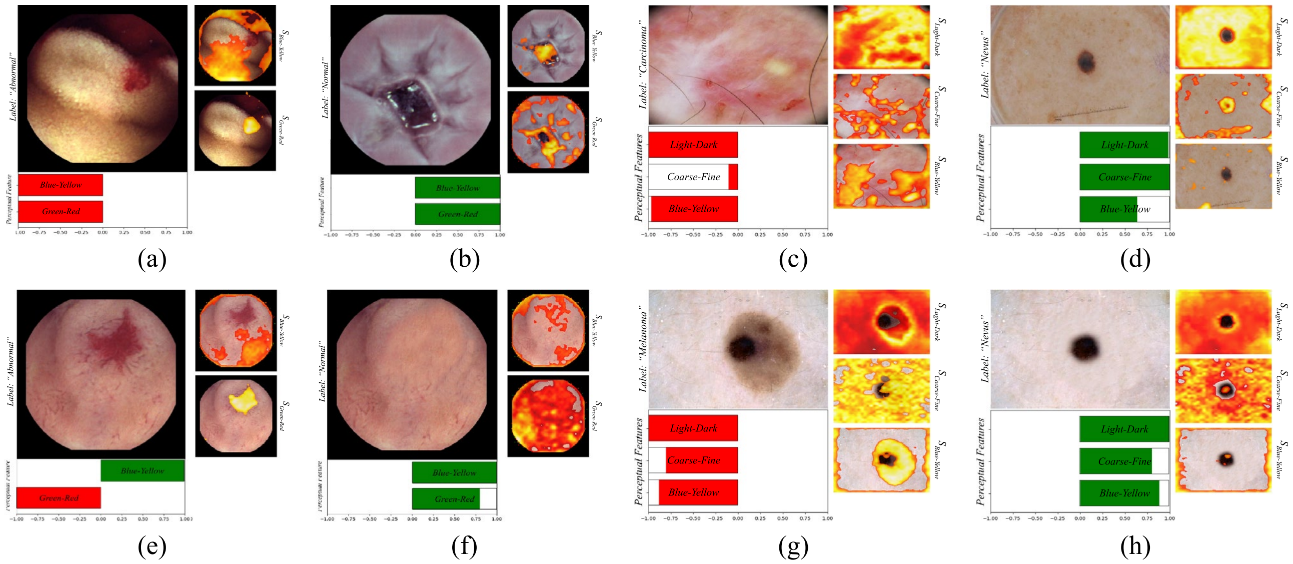
(b) When the shape resembles a banana, the *coarse–fine* PFM strongly suggests that the image belongs to the banana class (Fig. 7i,l). The magnitude of *light–dark* RSS has also changed, but still trends towards the apple class. The PRMs $S_{light\text{-}dark}$ and $S_{coarse\text{-}fine}$ appear to contribute to the segregation the depicted object; however, only the RSS of *coarse–fine* PFM suggests the opposite class, indicating that is more sensitive to shape variations.

(c) When the yellow region is added, the *light–dark* and *green-yellow* PFMs guide the prediction to the banana class. However, the *green–red* RSS trends towards the apple class. As expected, $S_{blue\text{-}yellow}$ and $S_{green\text{-}red}$ focus on the yellow and red segments of the object respectively (Fig. 7l). This justifies the trend of each PFM towards either class, i.e., yellow and red are representative colors of banana and apple class respectively.

These interpretations reveal that the *coarse–fine* PFM enables the respective sub-network to respond to different shape variations and infer relevant decisions. In addition, the color related PFMs, *i.e.*, *blue-yellow* and *green–red*, are very sensitive to the class-related colors and it is clearly reflected both in the respective PRMs and RSSs. When both the class-related colors, *i.e.*, *yellow* and *red*, cooccur in the image, the *blue-yellow* and *green–red* PFM guide the prediction towards the banana and apple class respectively.

Endoscopic datasets. The experiments showed that EPU_{II} tend to discriminate normal from abnormal images of the endoscopic datasets mainly based on the *blue-yellow* and *green–red* PFMs. This is illustrated in the respective global bar-charts (Fig. 6b–d). As it can be observed, the *light–dark* and *coarse–fine* are biased towards a specific class, in all endoscopic datasets. On the other hand, the chromatic PFMs are the main contributors to the correct classification predictions. This finding is in accordance with the literature since it has been proven that color has a leading role in finding abnormalities in the gastrointestinal tract[51], and with the results of the ablation study in Table 3.

An example of local bar-charts visualizing the prediction interpretations of EPU-CNN on endoscopic images is presented in Fig. 8a,b. Since the *light–dark* and *coarse–fine* features are not informative, only the color related PFMs were considered. Figures 8a,b illustrate correctly classified endoscopic images of both the normal and abnormal class. In the case of the image depicting an abnormality (Fig. 8a), the $S_{green\text{-}red}$ indicates that the focus of the sub-network that corresponds to the *green–red* PFM focuses on the abnormality, i.e., blood, whereas $S_{blue\text{-}yellow}$ focuses on normal tissue and only partially on the abnormal region.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 8. Example of EPU-CNN interpretations, as generated by EPU_{II}, on biomedical images. The label field indicates the predicted label. (**a**) Abnormal and (**b**) normal endoscopic image; (**c**) Carcinoma and (**d**) (normal) nevus skin lesion; (**e**) Abnormal endoscopic image and (**f**) modification of (**e**) to resemble a normal endoscopic image; (**g**) Melanoma skin lesion and (**h**) modification of (**g**) to resemble nevus.

To assess the behavior of EPU_{II} in a more controlled way in the endoscopic datasets, we proceeded to digitally process an endoscopic image and create different conditions for their classification to the normal and the abnormal classes. Indicative examples are presented in Fig. 8 where the abnormal region of Fig. 8e is removed, resulting in the synthetic image of Fig. 8f. The qualitative result of this process, considering only the chromatic PFMs, is reflected in the RSSs and the PRMs of Fig. 8f. Specifically, it can be noticed that by replacing the abnormal region with normal tissue, the RSS of the *green–red* PFM shifts from trending towards the abnormal class (red) to the normal class (green). Furthermore, the RSS of the *green–red* PFM, in the absence of an abnormality, indicate that the respective subnetwork focuses on normal tissue.

Dermoscopic datasets. The evaluation of the interpretability of EPU_{II} on the dermoscopic datasets revealed that all the PFMs participate actively in the classification process with an exception to *green–red* PFM that appears biased towards either the Melanoma or Carcinoma class on all trials (Fig. 6e–g). This is an indication that the PFM of *green–red* is not informative to the network regarding the classification of dermoscopic images. Furthermore, as it can be observed in Fig. 6e–g the classification process of EPU_{II} is relying on both chromatic and textural cues (i.e., *blue-yellow, light–dark* and *coarse–fine*) that are considered by the ABCD rule of skin lesion classification to assess the malignancy of a lesion[68] (this can also be confirmed from the results of the ablation study in Table 3).

An example of local bar-charts of classified dermoscopic images are illustrated in Fig. 8c,d. The local bar-chart includes the most informative PFMs, i.e., *light–dark*, *coarse–fine* and *blue-yellow*. In Fig. 8c,d all RSSs are trending, correctly, towards the abnormal (carcinoma, red) and normal (nevus, green) class respectively. In the case of the carcinoma (Fig. 8c), $S_{light\text{-}dark}$ focuses on the entirety of the image, whereas $S_{coarse\text{-}fine}$ and $S_{blue\text{-}yellow}$ focuses on regions with color variations, *e.g.*, on the yellow spot and little cuts on the lest and bottom side of the image respectively. In the case of the nevus (Fig. 8d), $S_{light\text{-}dark}$ and $S_{coarse\text{-}fine}$ isolate the lesion by segregating it from the rest of the image, either by focusing on it or around it, whereas $S_{blue\text{-}yellow}$ indicates only a slight attention of the network to the lesion. Similarly, to the other datasets, we proceeded to digitally modify the image of Fig. 8g that illustrates a melanoma to resemble a nevus. The modification was implemented according to the rule-based diagnostic criteria expressed by the ABCD rule[68]; in detail, we removed the part of the lesion that introduced color variation on the same mole and obtained a more symmetrical shape. The qualitative results of this process are illustrated in Fig. 8h, where it is shown that after the modification all the RSSs trend towards the nevus class (green). Furthermore, the $S_{blue\text{-}yellow}$ PRM, in the absence of an abnormal region, indicate that the respective subnetwork does not focus on the skin lesion. The $S_{light\text{-}dark}$ and $S_{coarse\text{-}fine}$ PRMs seem to maintain a similar behavior with the unmodified image.

Qualitative comparison with state-of-the-art interpretable methods.

Even though there is an increasing research interest regarding the interpretation of CNNs, there is still not a standard procedure to evaluate and compare the interpretable output. Nevertheless, a qualitative comparison can reveal strengths and weaknesses of such methods. In this study, the interpretations that EPU-CNN provides are qualitatively compared to seven methodologies that have been proposed to interpret CNNs and have been also widely used in the literature. These methods provide saliency maps indicating regions or points on the input image that are estimated to be crucial for a prediction inferred by a CNN.

In detail, six *post-hoc* methodologies, namely, Grad-CAM[69], LIME[8], XRAI[70], Shapley Additive exPlanations (SHAP)[71], Smoothgrad[72] and Vanilla Gradients[73], as well as one inherently interpretable model[65] (TT) were utilized in this evaluation. The *post-hoc* methodologies were applied on the CNN models that achieved the highest performance on each dataset according to Table 1, whereas TT was trained on each dataset from scratch. All the methods provide interpretations in the form of saliency maps while TT can also provide bounding boxes that specify discreetly the estimated region of interest. These methods were selected since they can render CNN models interpretable without the need for training on datasets specifically annotated for interpretable learning, *e.g.*, with annotation regarding the concepts that are depicted on images.

Figure 9 summarizes the interpretations provided by each method on exemplary images that are presented in Figs. 7 and 8. All the images have been correctly classified by the respective models that were used. In detail, only XRAI and SHAP were successful at highlighting regions of interest on the images that can be regarded crucial for classification, i.e., areas of the apple, the skin lesion and blood depicted in the endoscopic image. The gradient-based interpretation approaches, i.e., Grad-CAM, Smoothgrad and Vanilla Grad, also revealed that the respective CNN models focus on image regions that can be regarded meaningful; nevertheless, the fuzziness of their visualization makes the communication of their interpretations difficult to comprehend. On the other hand, EPU-CNN can provide different visualizations, that highlight the most relevant regions with respect to each PFM as it was estimated by the layer activations of each subnetwork. This can also be expressed quantitatively since the RSSs indicate the degree to which each highlighted region affects the classification result. Furthermore, the dataset-wide plots that can be constructed by using an EPU-CNN model give insights regarding which PFMs are important for classifying the images of a particular dataset. To the best of our knowledge no other interpretation approach can incorporate all this information to its explanations and simultaneously be applied on non-specialized datasets, *e.g.*, datasets where each image is annotated only with respect to their class membership.

To demonstrate the applicability of EPU-CNN framework, the best EPU-CNN model in terms of classification performance on the previous experiments, i.e., EPU_{II}, was trained and tested on four non-biomedical benchmark datasets including CIFAR-10, MNIST, fashion MNIST and iBean datasets. To train EPU_{II} in a multiclass setting the SoftMax activation function replaced the final log-sigmoid $\boldsymbol{\sigma}(\cdot)$ (Eq. 5) function. Figure 10 illustrates interpretations provided by EPU_{II} on predictions of images on the CIFAR-10 dataset, and Fig. 11 illustrates indicative interpretations obtained for the classification of images from the MNIST, Fashion MNIST, and iBean

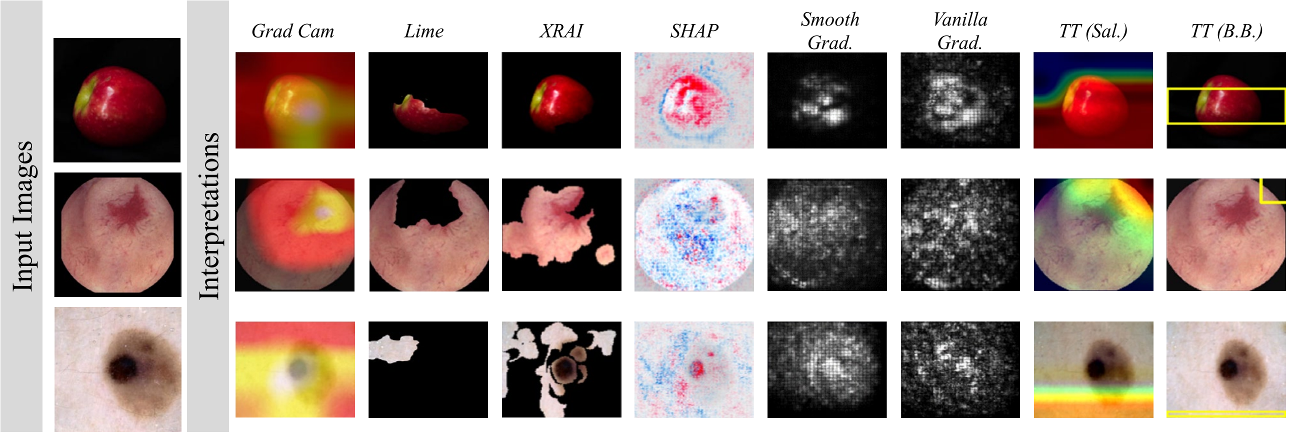
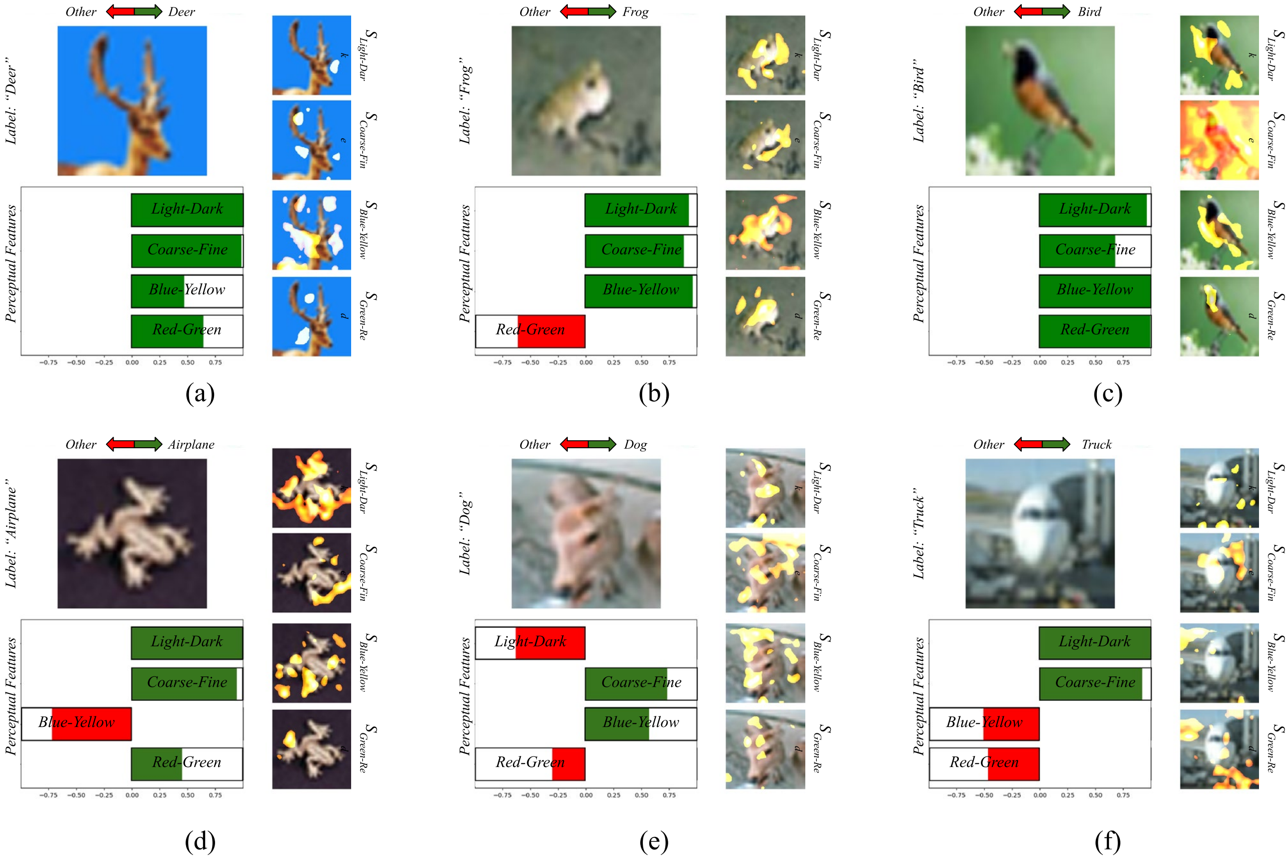


Figure 9. Example of CNN interpretations provided by various methodologies.



(a) (b) (c)

(d) (e) (f)

Figure 10. Example of EPU-CNN interpretations, as generated by EPU_{II}, on images of the CIFAR-10 dataset. The label field indicates the predicted label. The first and second rows illustrate interpretations of correct and wrong predictions, respectively.

datasets. For consistency with the interpretations of the binary settings, the respective illustrations depict how each PFM drives a prediction towards either the predicted (green) or any other class (red). The first and second rows of Figs. 10 and 11 illustrate interpretations of correct and wrong classifications on images included in the aforementioned datasets, respectively. As it can be observed the PRMs generated by the EPU_{II} on the interpretations presented in the first row of Fig. 10 highlight the object of interest with more precision when compared to the wrongly classified images in the second row of Fig. 10. For example, in Fig. 10b all the PRMs highlight regions of the frog, and the image has been correctly classified. On the other hand, in Fig. 10d the PRMs mainly highlight regions around the frog and the respective image is misclassified to the airplane class, based on all the PFMs except from the *blue–yellow*. Furthermore, it can be noticed that the respective RSSs behave similarly, i.e., in Fig. 10c the PRM of *coarse–fine* highlights the whole image and it does not focus solely on the bird. Accordingly,

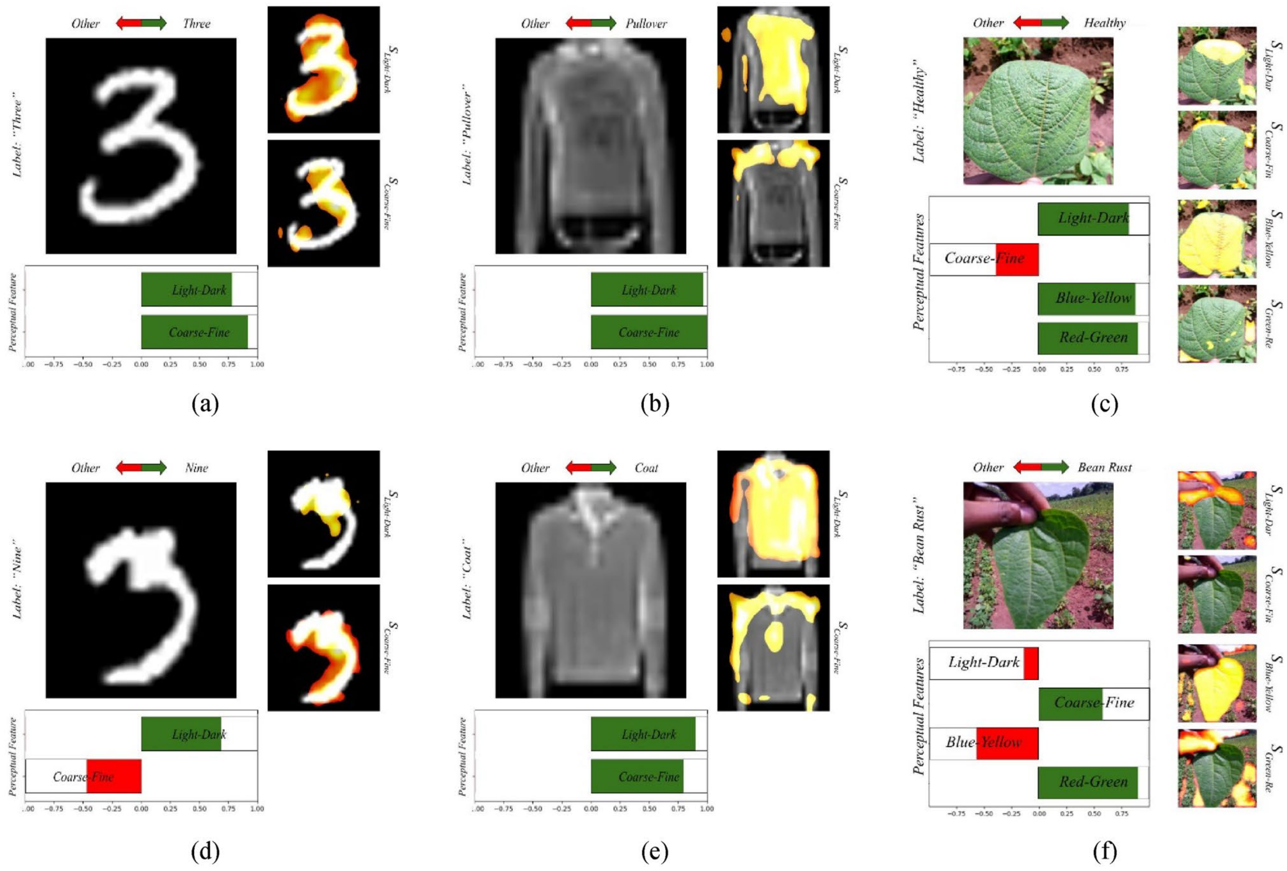


Figure 11. Example of EPU-CNN interpretations, as generated by EPU_{II}, on images of the MNIST, fashion MNIST and iBean datasets. The label field indicates the predicted label. The first and second rows illustrate interpretations of correct and wrong predictions, respectively.

the respective RSS of *coarse–fine* has a smaller magnitude towards the correct class than the rest of RSSs, which, based on the respective PRMs consider mainly the region of the bird.

Moreover, in Fig. 10e the image that belongs to the deer class is misclassified as a dog. As it can be noticed, the PRMs of *coarse–fine* and *blue–yellow* are mainly focusing on the head region of the deer and the respective RSSs drives the prediction towards the dog class. This can be attributed to the fact that the particular deer does not seem to have horns, and it has a color pattern that matches that of the dog class. This can be further substantiated by observing the image of Fig. 10a. This image depicts a deer that has been correctly classified by the EPU-CNN model. As it can be noticed, the PRMs of *coarse–fine* and *blue–yellow* highlight the head region of the deer where the horns are present. Finally, the image of Fig. 10f that depicts an airplane is classified to the truck class based on the PFMs of *light–dark* and *coarse–fine*. As it can be noticed, all the PRMs focus on the body of the airplane, and on the wheels, whereas no PRM focuses on the wings. Therefore, the respective RSSs of *light–dark* and *coarse–fine* drive the prediction with a higher magnitude towards the truck class. Figure 12a presents a comparison in terms of classification accuracy among EPU_{II} (orange bar) and other state-of-the-art CNN models[62–64,74,75] (gray bars) on the CIFAR-10 dataset. EPU_{II} achieved an accuracy score of 93.31% which is comparable or better than the other models considered. However, a major advantage over the other models is that the EPU-CNN model can provide interpretations regarding the classification outcome. Similarly, Fig. 12b presents a comparison of the classification performance in terms of accuracy among EPU_{II} and other CNN models that have been proposed for classifying images of the iBean dataset[76]. The classification results of EPU_{II} compared to other CNN models[77] on the MNIST and Fashion MNIST datasets are presented in Fig. 12c,d, respectively. EPU_{II} provided an accuracy of 92.32% outperforming the other models on iBean dataset. Regarding the MNIST and Fashion MNIST datasets, EPU_{II} achieved a classification accuracy of 99.44% and 93.20%, respectively, which is comparable to the other models considered. In the case of the well-known CIFAR-10, MNIST and Fashion MNIST datasets the regularization effect of early stopping was performed to prevent overfitting on the training set, considering the size of data and the model size[78].

To further demonstrate the interpretations of EPU-CNN, as it can be observed in Fig. 11, the depicted interpretations for each dataset display the same object class, classified correctly in the first row of Fig. 11 and classified wrongly in the second row of Fig. 11, respectively. Specifically, for the MNIST dataset in Fig. 11a both PRMs (*light–dark* and *coarse–fine*) mainly highlight the part of the image depicting the handwritten digit for the correct classification (both RSSs are positive), whereas in Fig. 11d the PRM of *light–dark* focuses on the brighter region of the image, which is not clearly defined making digit '3' to resemble digit '9'. Thus, the respective image

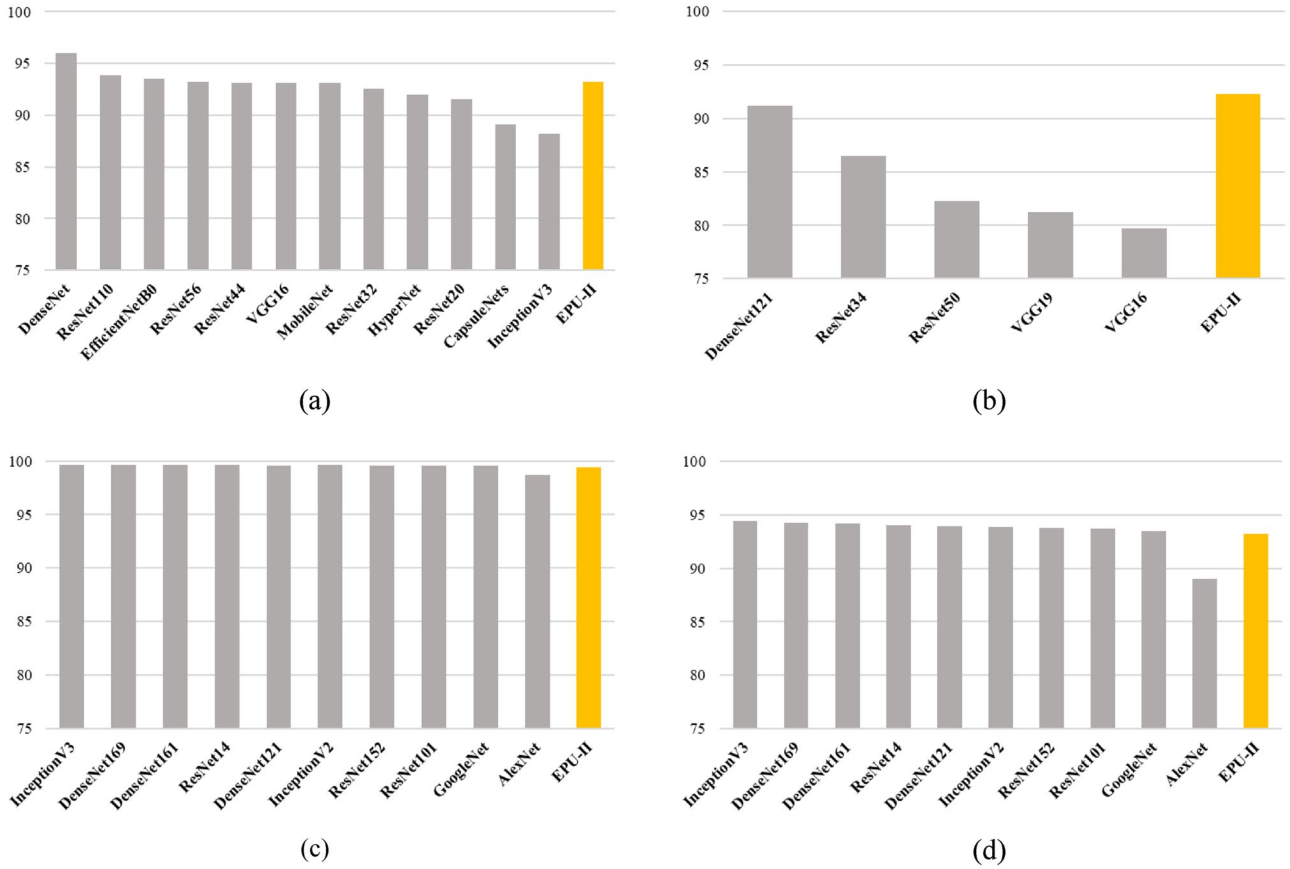


Figure 12. Classification performance in terms of accuracy on the (**a**) CIFAR-10 dataset, (**b**) iBean dataset, (**c**) MNIST dataset and (**d**) on Fashion MNIST dataset.

is misclassified as digit '9'. This is also justified based on the *coarse–fine* RSS, which is negative for that case, explaining quantitatively that digit '3' resembles more to a digit of another class, different than digit '9', which is the predicted one. Additionally, for the fashion MNIST dataset, in both cases illustrated in the second column of Fig. 11 the respective PRMs of *coarse–fine* mainly focus on the upper part of the images, where the neckline of the fashion item is depicted. The PRMs of *light–dark* mainly highlight a broader region of the fashion items. In Fig. 11b both RSSs are positive, explaining that the image indeed resembles a pullover, which is the correctly predicted class. Figure 11e depicts a pullover, which is misclassified as a coat. Both RSSs show that the prediction tends towards the coat class, which is incorrect. This is also justified by the coarse–fine PRM, which highlights the neckline and buttons, since they are more likely to appear on a coat, rather than a pullover. In the example image from the iBean dataset, the healthy leaf image in Fig. 11f is misclassified as bean rust, based on RSSs of red-green and coarse–fine, which can be attributed to the color patterns highlighted by the respective PRMs in the background of the image, which are less likely to be considered healthy leaf patterns. We can also observe a similar behavior of driving the prediction to the wrong direction of the coarse–fine RSS in the case of the correct classification in Fig. 11c, where the PFM captures parts of the background and drives the prediction towards a non-healthy class.

Quantitative assessment of the perceptual relevance maps.

To further assess the PRMs in terms of their capacity to indicate image regions that are relevant for the classification of images by an EPU-CNN model, the Remove and Debias[79] (ROAD) metric has been used. The ROAD metric measures how a given saliency map, generated by a gradient-based method, influences the classification confidence of the network. A saliency map indicates regions on the input image that contribute to a certain classification result. ROAD examines how the confidence of the network changes when spatial regions of the image that are indicated by the saliency map are modified. By identifying regions of low and high importance, according to an interpretation saliency map at a specific threshold, ROAD isolates the respective image regions of an image I and imputes the remaining image values. This results in two new perturbated images referred to as R_{low} and R_{high}, each of which isolates the least and highest important information, respectively. These images are propagated sequentially to a classification model f and the change in the classification confidence, cc, is assessed as follows:

$$cc_{low} = f(R_{low}) - f(I), \tag{10}$$

$$cc_{high} = f\left(R_{high}\right) - f(I). \tag{11}$$

Once the change in the classification confidence is estimated by considering both R_{low} and R_{high}, the final ROAD score is estimated as:

$$ROAD = \frac{cc_{low} - cc_{high}}{2}. \tag{12}$$

Higher ROAD scores indicate more accurate interpretations.

To evaluate the interpretation capacity of the proposed PRM estimation approach, the best performing EPU-CNN, i.e., EPU_{II}, was selected. The ROAD metric was calculated by considering all the PRMs that are estimated by each EPU-CNN prediction. For comparison purposes the proposed PRM estimation method that EPU_{II} utilized was replaced by GradCAM. For one-to-one evaluation, the GradCAM methodology was applied on the convolutional layers that were selected (by ablation study) to be more informative for the PRM estimation, i.e., the 5th convolutional layer of each subnet. Additionally, a recent ensemble model used along with post-hoc methods for interpretable classification was considered in the comparative evaluation[33]. This model was composed of four base models, namely a VGG-16, a DenseNet, an Xception and a ResNet model. GradCam was used as a post-hoc approach to make this ensemble model interpretable by aggregating the saliency maps estimated by its application on each model. We refer to the GradCam EPU-CNN and to the ensemble model as $EPU_{GradCam}$ and *Ensemble*, respectively.

For a more thorough investigation of the generalization capabilities of EPU-CNN, the ROAD metric was calculated on the models trained on MNIST, Fashion MNIST, CIFAR-10, iBean, Banapple as well as on the Endoscopic and Dermoscopic datasets. Table 4 summarizes the average ROAD scores achieved by EPU_{II}, $EPU_{GradCAM}$ and *Ensemble* models. As it can be observed, the results obtained using the proposed methodology of the PRM provided by EPU_{II} model outperform the application of GradCAM in all datasets.

Conclusions and future work

In this study we proposed a novel, generalized framework, called EPU-CNN, that provides a guideline for the development of interpretable CNN models, inspired by GAMs. A model, designed according to EPU-CNN framework, consists of an ensemble of sub-networks with a base CNN architecture that is trained as one. The proposed framework can be used to render conventional CNN models interpretable, by using it as a base model. The proposed EPU-CNN framework can be constructed as an ensemble of an arbitrary number of base CNN sub-networks, given a respective number of PFMs, feeding each sub-network. PFMs should satisfy the following requirements: (a) they should represent perceptually relevant opponent features of the images, and (b) they should be orthogonal between each other. In this paper a total of four PFMs were proposed for interpretable classification of images based on color and texture, which are two decisive properties for image understanding. These PFMs were chosen according to the literature of cognitive science and human perception. Considering that these PFMs are based on the representation of visual information in the human visual system, they are sufficiently generic, suitable for a wide variety of applications. The generality of EPU-CNN framework enables its use with different input PFMs (satisfying the above requirements), that could be derived using other transformations resulting in opponent image components with a physical interpretation perceivable by humans. EPU-CNN is designed in a way enabling human-friendly interpretations of its classification results based on the utilized perceptual features. The interpretations provided by EPU-CNN are in the form of RSSs that quantify the resemblance of a perceptual feature to a respective class. These interpretations are complemented by PRMs indicating the image regions where the network focuses to infer its interpretable decisions. Furthermore, EPU-CNN provides spatial expression of an explanation on the input image. The most important conclusions of this study can be summarized as follows:

- EPU-CNN models satisfy the need for interpretable models based on human perception, i.e., the proposed framework is able to provide interpretations in accordance with human perception and cognitive science, e.g., EPU-CNN classifies endoscopic images based on the chromatic perceptual features.
- Unlike other inherently interpretable CNN methodologies[21,42], the classification performance of EPU-CNN models is not affected by their capacity to provide interpretations. In fact, the results obtained from the comparison of EPU-CNN models with respective non-interpretable CNN models, show that their performance is better or at least comparable to that of the non-interpretable models.
- When an image is modified with respect to a perceptual feature, e.g., color, the interpretations derived from the EPU-CNN model change accordingly both on natural and biomedical images (Figs. 6, 7).

Models	**ROAD**						
	MNIST	***Fashion MNIST***	***Cifar-10***	***iBean***	***Endoscopic***	***Dermoscopic***	***Banapple***
EPU_{II}	0.06	0.06	0.11	0.15	0.15	0.16	0.09
$EPU_{GradCam}$	0.02	0.02	0.03	0.11	−0.07	0.06	0.07
Ensemble	−0.01	0.01	0.03	0.02	0.01	0.03	−0.02

Table 4. Average ROAD scores obtained from different models on different datasets.

- Since EPU-CNN is a generalized framework, it provides a template for the development of interpretable CNNs that fulfill the requirements imposed by current legislations regarding the commercial applicability of ML models.
- EPU-CNN can be particularly useful in the context of biomedical applications, considering that abnormalities are usually characterized by color and texture differentiations of human tissues, as indicated by the results obtained from its application on endoscopic and dermoscopic datasets.

Creating a perceptually interpretable CNN-based image classification system requires that, somehow, knowledge about the way humans perceive images is introduced into that system. The initial decomposition of the input images into different PFMs that provide separate information about color and texture components of the images, serves this purpose. The manual definition of PFMs could be considered as a limitation of the proposed framework; however, it should be considered more as a way of modeling human perception into an interpretable classification system, rather than a way to introduce an application-dependent bias into the feature extraction process, as it is typically done with the traditional, so-called "hand-engineered" or "hand-crafted" feature extraction. Traditional feature extraction involves the calculation of specific features from the images, whereas EPU-CNN maintains its ability to extract automatically derived features from the images, by applying this process on the PFMs, which are also images, representing complementary information of the input images. Most inherently interpretable models that have been proposed in the literature, can only be applied on datasets that are further annotated with respect to human-understandable concepts illustrated in each image, which results in limitation regarding their applicability[19,20]. The PFM selection of an EPU-CNN model can be considered as a less demanding and time-consuming procedure when compared to the annotation of huge datasets with the human-understandable concepts.

This paper also showed that given an EPU-CNN configuration with a defined set of PFMs, it is possible to use global bar-charts (revealing the overall contribution of PFMs to the data discrimination process) to select a subset of PFMs that are more relevant to a particular application. This is optional but it can reduce the complexity of the EPU-CNN, since the respective sub-networks will be pruned. Furthermore, it can offer focused interpretations and insights based on the most relevant features regarding the internal process of the EPU-CNN model. For example, in the case of endoscopic images, the EPU-CNN models considered only the PFMs of color as more important which is in accordance with the respective literature; thus, only two of the four sub-networks were sufficient. In the case of greyscale input images, such as the images of the MNIST dataset, the PFMs representing color can automatically be discarded considering their file format.

Future work includes investigation of alternative PFM representations based on transformation other than CIE-*Lab* and 2D DWT considered in this paper, and applications on different domains. User evaluation studies with domain experts could contribute in further adaptation of the proposed framework to derive more meaningful interpretations considering the domain semantics, e.g., in endoscopy associating colors with pathologies. Another perspective is further automation of the PFM definition and subset selection processes towards a direction that minimizes human intervention and is more compatible with the principles of deep learning.

Data availability

Banapple is available at https://github.com/innoisys/Banapple. KID is available at: https://mdss.uth.gr/datasets/endoscopy/kid/. Kvasir is available at: https://datasets.simula.no/kvasir/. ISIC2019 is available at: https://challenge2019.isic-archive.com/. CIFAR-10 is available at: https://www.cs.toronto.edu/~kriz/cifar.html. MNIST is available at: http://yann.lecun.com/exdb/mnist/. Fashion MNIST is available at: https://github.com/zalandoresearch/fashion-mnist. iBean dataset is available at: https://github.com/AI-Lab-Makerere/ibean/.

Code availability

Code is available at https://github.com/innoisys/EPU-CNN.

Received: 18 November 2022; Accepted: 8 July 2023
Published online: 14 July 2023

References

1. Selbst, A. & Powles, J. Meaningful information and the right to explanation. In *Conference on Fairness, Accountability and Transparency* 48–48 (2018).
2. Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I. & Atkinson, P. M. Explainable artificial intelligence: An analytical review. *Wiley Interdiscipl. Rev. Data Mining Knowl. Discov.* **11**, e1424 (2021).
3. Rudin, C. *et al.* Interpretable machine learning: Fundamental principles and 10 grand challenges. *Stat. Surv.* **16**, 1–85 (2022).
4. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci.* **116**, 22071–22080 (2019).
5. Muñoz-Romero, S., Gorostiaga, A., Soguero-Ruiz, C., Mora-Jiménez, I. & Rojo-Álvarez, J. L. Informative variable identifier: Expanding interpretability in feature selection. *Pattern Recogn.* **98**, 107077 (2020).
6. Yao, K., Cao, F., Leung, Y. & Liang, J. Deep neural network compression through interpretability-based filter pruning. *Pattern Recogn.* **119**, 108056 (2021).
7. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. Preprint at http://arXiv.org/1705.07874 (2017).
8. Ribeiro, M. T., Singh, S. & Guestrin, C. Why should I trust you? Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (2016).
9. Mittelstadt, B., Russell, C. & Wachter, S. Explaining explanations in AI. In *Proc. Conference on Fairness, Accountability, and Transparency* 279–288 (2019).
10. Yu, L., Xiang, W., Fang, J., Chen, Y.-P.P. & Zhu, R. A novel explainable neural network for Alzheimer's disease diagnosis. *Pattern Recogn.* **131**, 108876 (2022).
11. Adebayo, J. *et al.* Sanity checks for saliency maps. Preprint at http://arXiv.org/1810.03292 (2018).

12. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
13. Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. Explainable AI: A review of machine learning interpretability methods. *Entropy* **23**, 18 (2020).
14. Lakkaraju, H., Bach, S. H. & Leskovec, J. Interpretable decision sets: A joint framework for description and prediction. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1675–1684 (2016).
15. Yang, G., Ye, Q. & Xia, J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Inf. Fusion* **77**, 29–52 (2022).
16. Sharma, A. & Mishra, P. K. Covid-MANet: Multi-task attention network for explainable diagnosis and severity assessment of COVID-19 from CXR images. *Pattern Recogn.* 31, 108826 (2022).
17. Chen, R., Chen, H., Ren, J., Huang, G. & Zhang, Q. Explaining neural networks semantically and quantitatively. In *Proc. IEEE/CVF International Conference on Computer Vision* 9187–9196 (2019).
18. Liang, H. *et al.* Training interpretable convolutional neural networks by differentiating class-specific filters. In *European Conference on Computer Vision* 622–638 (2020).
19. Bau, D., Zhou, B., Khosla, A., Oliva, A. & Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 6541–6549 (2017).
20. Barbiero, P. *et al.* Entropy-based logic explanations of neural networks. *Proc. AAAI Conf. Artif. Intell.* **36**, 6046–6054 (2022).
21. Zhang, Q., Yang, Y., Ma, H. & Wu, Y. N. Interpreting cnns via decision trees. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 6261–6270 (2019).
22. Hastie, T. J. & Tibshirani, R. J. *Generalized Additive Models* Vol. 43 (CRC Press, 1990).
23. Arrieta, A. B. *et al.* Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020).
24. Cai, Y., Zheng, J., Zhang, X., Jiang, H. & Huang, M.-C. GAM feature selection to discover predominant factors for mortality of weekend and weekday admission to the ICUs. *Smart Health* **18**, 100145 (2020).
25. Yang, Z., Zhang, A. & Sudjianto, A. GAMI-Net: An explainable neural network based on generalized additive models with structured interactions. *Pattern Recogn.* **120**, 108192 (2021).
26. Jung, J. H. & Kwon, Y. Color, edge, and pixel-wise explanation of predictions based on interpretable neural network model. In *2020 25th International Conference on Pattern Recognition (ICPR)* 6003–6010 (2021).
27. Greisdorf, H. & O'Connor, B. Modelling what users see when they look at images: A cognitive viewpoint. *J. Document.* 58, 6-29 (2002).
28. Wolpert, D. H. Stacked generalization. *Neural Netw.* **5**, 241–259 (1992).
29. Xiang, A. & Wang, F. Towards interpretable skin lesion classification with deep learning models. In *AMIA Annual Symposium Proceedings*, Vol. 2019, 1246 (2019).
30. Shorfuzzaman, M. An explainable stacked ensemble of deep learning models for improved melanoma skin cancer detection. *Multimedia Syst.* **28**, 1309–1323 (2022).
31. Bany Muhammad, M. & Yeasin, M. Interpretable and parameter optimized ensemble model for knee osteoarthritis assessment using radiographs. *Sci. Rep.* **11**, 14348 (2021).
32. Liz, H. *et al.* Ensembles of convolutional neural network models for pediatric pneumonia diagnosis. *Futur. Gener. Comput. Syst.* **122**, 220–233 (2021).
33. Alfi, I. A., Rahman, M. M., Shorfuzzaman, M. & Nazir, A. A non-invasive interpretable diagnosis of melanoma skin cancer using deep learning and ensemble stacking of machine learning models. *Diagnostics* **12**, 726 (2022).
34. Tanaka, J., Weiskopf, D. & Williams, P. The role of color in high-level vision. *Trends Cogn. Sci.* **5**, 211–215 (2001).
35. Friconnet, G. Exploring the correlation between semantic descriptors and texture analysis features in brain MRI. *Chin. J. Acad. Radiol.* **4**, 105–115 (2021).
36. Oukil, S., Kasmi, R., Mokrani, K. & García-Zapirain, B. Automatic segmentation and melanoma detection based on color and texture features in dermoscopic images. *Skin Res. Technol.* **28**, 203–211 (2022).
37. Rauf, H. T. *et al.* A citrus fruits and leaves dataset for detection and classification of citrus diseases through machine learning. *Data Brief* **26**, 104340 (2019).
38. Khaled, A. Y., Parrish, C. A. & Adedeji, A. Emerging nondestructive approaches for meat quality and safety evaluation—A review. *Comp. Rev. Food Sci. Food Saf.* **20**, 3438–3463 (2021).
39. Yang, J., Wang, C., Jiang, B., Song, H. & Meng, Q. Visual perception enabled industry intelligence: State of the art, challenges and prospects. *IEEE Trans. Ind. Inf.* **17**, 2204–2219 (2020).
40. Kondratyuk, D., Tan, M., Brown, M. A. & Gong, B. When ensembling smaller models is more efficient than single large models. Preprint at http://arXiv.org/2005.00570 (2020).
41. Ganaie, M. A., Hu, M., Malik, A., Tanveer, M. & Suganthan, P. Ensemble deep learning: A review. *Eng. Appl. Artif. Intell.* **115**, 105151 (2022).
42. Zhang, Q., Wu, Y. N. & Zhu, S.-C. Interpretable convolutional neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 8827–8836 (2018).
43. Deroy, O. Object-sensitivity versus cognitive penetrability of perception. *Philos. Stud.* **162**, 87–107 (2013).
44. Woniak, M., Grana, M. & Corchado, E. A survey of multiple classifier systems as hybrid systems. *Inf. Fusion* **16**, 3–17 (2014).
45. Hurvich, L. M. & Jameson, D. An opponent-process theory of color vision. *Psychol. Rev.* **64**, 384 (1957).
46. Chatterjee, S. & Callaway, E. M. Parallel colour-opponent pathways to primary visual cortex. *Nature* **426**, 668–671 (2003).
47. Poirson, A. B. & Wandell, B. A. Pattern—Color separable pathways predict sensitivity to simple colored patterns. *Vis. Res.* **36**, 515–526 (1996).
48. Mäenpää, T. & Pietikäinen, M. Classification with color and texture: Jointly or separately? *Pattern Recogn.* **37**, 1629–1640 (2004).
49. Hansen, T. & Gegenfurtner, K. R. Independence of color and luminance edges in natural scenes. *Vis. Neurosci.* **26**, 35–49 (2009).
50. Wyszecki, G. & Stiles, W. S. *Color Science: Concepts and Methods, Quantitative Data and Formulae* 2nd edn. (Wiley-Interscience, 2000).
51. Iakovidis, D. K. & Koulaouzidis, A. Automatic lesion detection in wireless capsule endoscopy—A simple solution for a complex problem. In *2014 IEEE International Conference on Image Processing (ICIP)* 2236–2240 (2014).
52. Huang, P.-W. & Dai, S. Image retrieval by texture similarity. *Pattern Recogn.* **36**, 665–679 (2003).
53. Mallat, S. G. A theory for multiresolution signal decomposition: The wavelet representation. In *Fundamental Papers in Wavelet Theory* 494–513 (2009).
54. Tuceryan, M. & Jain, A. K. Texture analysis. In *Handbook of Pattern Recognition and Computer Vision* 235–276 (1993).
55. Biederman, I. & Ju, G. Surface versus edge-based determinants of visual recognition. *Cogn. Psychol.* **20**, 38–64 (1988).
56. Iakovidis, D. K., Georgakopoulos, S. V., Vasilakakis, M., Koulaouzidis, A. & Plagianakos, V. P. Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification. *IEEE Trans. Med. Imaging* **37**, 2196–2210 (2018).
57. Yen, J.-C., Chang, F.-J. & Chang, S. A new criterion for automatic multilevel thresholding. *IEEE Trans. Image Process.* **4**, 370–378 (1995).
58. Koulaouzidis, A. *et al.* KID Project: An internet-based digital video atlas of capsule endoscopy for research purposes. *Endosc. Int. Open* **5**, E477 (2017).

59. Pogorelov, K. *et al.* KVASIR: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proc. 8th ACM on Multimedia Systems Conference* 164–169. https://doi.org/10.1145/3083187.3083212 (2017).
60. Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III*, Vol. 9351 (Springer, 2015).
61. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer statistics, 2021. *CA Cancer J. Clin.* **71**, 7–33 (2021).
62. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. Preprint at http://arXiv.org/1409.1556 (2014).
63. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016).
64. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 4700–4708 (2017).
65. Chen, C. *et al.* This looks like that: Deep learning for interpretable image recognition. *Adv. Neural Inf. Process. Syst.* **32**, 1 (2019).
66. Provost, F. & Fawcett, T. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions In *Proc. 3rd International Conference on Knowledge Discovery and Data Mining* (1997).
67. Jaccard, P. The distribution of the flora in the alpine zone. 1. *New Phytol.* **11**, 37–50 (1912).
68. Nachbar, F. *et al.* The ABCD rule of dermatoscopy: High prospective value in the diagnosis of doubtful melanocytic skin lesions. *J. Am. Acad. Dermatol.* **30**, 551–559 (1994).
69. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. IEEE International Conference on Computer Vision* 618–626 (2017).
70. Kapishnikov, A., Bolukbasi, T., Viégas, F. & Terry, M. Xrai: Better attributions through regions. In *Proc. IEEE/CVF International Conference on Computer Vision* 4948–4957 (2019).
71. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural. Inf. Process. Syst.* **30**, 4765–4774 (2017).
72. Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. Smoothgrad: Removing noise by adding noise. Preprint at http://arXiv.org/1706.03825 (2017).
73. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. Preprint at http://arXiv.org/1312.6034 (2013).
74. Ha, D., Dai, A. & Le, Q. V. Hypernetworks. Preprint at http://arXiv.org/1609.09106 (2016).
75. Sabour, S., Frosst, N. & Hinton, G. E. Dynamic routing between capsules. *Adv. Neural Inf. Process. Syst.* **30**, 1 (2017).
76. Abed, S. H., Al-Waisy, A. S., Mohammed, H. J. & Al-Fahdawi, S. A modern deep learning framework in robot vision for automated bean leaves diseases detection. *Int. J. Intell. Robot. Appl.* **5**, 235–251 (2021).
77. Gavrikov, P. & Keuper, J. Cnn filter db: An empirical investigation of trained convolutional filters. *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 19066–19076 (2022).
78. Shen, R., Gao, L. & Ma, Y.-A. On optimal early stopping: Over-informative versus under-informative parametrization. Preprint at http://arXiv.org/2202.09885 (2022).
79. Rong, Y., Leemann, T., Borisov, V., Kasneci, G. & Kasneci, E. A consistent and efficient evaluation strategy for attribution methods. Preprint at http://arXiv.org/2202.00449 (2022).

Acknowledgements

The authors acknowledge support of this work by the project "Smart Tourist" (MIS 5047243) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

Author contributions

G.D., D.K.I. and E.C. conceived the study. G.D. and E.C. developed methods, designed visualizations and metrics, ran experiments, and contributed to the writing. D.K.I. supervised research, method development, experiments and contributed to the writing.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.K.I.