Universidade Estadual de Campinas
Instituto de Computação

Rosa Yuliana Gabriela Paccotacya Yanque

A Comparative Analysis of eXplainable Artificial Intelligence Methods for Skin Lesion Classification

Uma Análise Comparativa de Métodos de Inteligência Artificial Explicáveis para Classificação de Lesões de Pele

CAMPINAS
2022

Rosa Yuliana Gabriela Paccotacya Yanque

A Comparative Analysis of eXplainable Artificial Intelligence Methods for Skin Lesion Classification

Uma Análise Comparativa de Métodos de Inteligência Artificial Explicáveis para Classificação de Lesões de Pele

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestra em Ciência da Computação.

Dissertation presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Computer Science.

Supervisor/Orientadora: Profa. Dra. Sandra Eliza Fontes de Avila

Este exemplar corresponde à versão final da Dissertação defendida por Rosa Yuliana Gabriela Paccotacya Yanque e orientada pela Profa. Dra. Sandra Eliza Fontes de Avila.

CAMPINAS
2022

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

P114c

Paccotacya Yanque, Rosa Yuliana Gabriela, 1996-
A comparative analysis of eXplainable artificial intelligence methods for skin lesion classification / Rosa Yuliana Gabriela Paccotacya Yanque. – Campinas, SP : [s.n.], 2022.

Orientador: Sandra Eliza Fontes de Avila.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Explicabilidade (Aprendizado de máquina). 2. Interpretabilidade (Aprendizado de máquina). 3. Melanoma. 4. Aprendizado profundo. 5. Imagens médicas. I. Avila, Sandra Eliza Fontes de, 1982-. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Informações Complementares

Título em outro idioma: Uma análise comparativa de métodos de inteligência artificial explicáveis para classificação de lesões de pele
Palavras-chave em inglês:
Explainability (Machine learning)
Interpretability (Machine learning)
Melanoma
Deep learning
Medical images
Área de concentração: Ciência da Computação
Titulação: Mestra em Ciência da Computação
Banca examinadora:
Sandra Eliza Fontes de Avila [Orientador]
Agma Juci Machado Traina
Emely Pujólli da Silva
Data de defesa: 20-12-2022
Programa de Pós-Graduação: Ciência da Computação

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-4352-7377
- Currículo Lattes do autor: http://lattes.cnpq.br/5789844358879056

Universidade Estadual de Campinas
Instituto de Computação

Rosa Yuliana Gabriela Paccotacya Yanque

A Comparative Analysis of eXplainable Artificial Intelligence Methods for Skin Lesion Classification

Uma Análise Comparativa de Métodos de Inteligência Artificial Explicáveis para Classificação de Lesões de Pele

Banca Examinadora:

- Profa. Dra. Sandra Eliza Fontes de Avila
  Universidade Estadual de Campinas
- Profa. Dra. Agma Juci Machado Traina
  Universidade de São Paulo
- Dra. Emely Pujólli da Silva
  Universidade Estadual de Campinas

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Campinas, 20 de dezembro de 2022

Only if we understand, will we care.
Only if we care, will we help.
Only if we help shall all be saved.

(Jane Goodall)

Acknowledgements

First, I would like to acknowledge and thank Professor Sandra Avila for her continuous support, motivation, understanding, and inspiration. I consider myself incredibly fortunate to have had her as my M.Sc. advisor and I am deeply grateful for her patience, encouragement, guidance, and belief in me.

I thank my colleagues from the Recod.ai laboratory and the friends I made in Brazil. They introduced Campinas to me and have been a source of emotional support, joy, encouragement, and unforgettable memories.

I am grateful to the University of Campinas and Recod.ai for providing me with the infrastructure, resources, and facilities necessary to conduct my research. This study was financed in part by The Brazilian National Council for Scientific and Technological Development (CNPq), grant #155459/2019-8. I would also like to express gratitude to the Bolsa Alumni from the Institute of Computing for their financial support during the last two months.

I am thankful to the LatinX in AI community for the mentoring programs and their various grants, which have allowed me to participate in cutting-edge AI conferences. I've been able to expand my knowledge and abilities as an AI researcher thanks to these opportunities.

I want to thank my mom, Nelly Yanque Ch., the strongest woman I've ever met and who embodies the saying "nothing is impossible". Her unconditional love and dedication have been a continual source of inspiration and motivation in my life. From my earliest memories, she's lighted the way for me to pursue my dreams with confidence. I am forever grateful for her unwavering support, example, and inspiring words.

I would like to express my gratitude to my father and my family, uncles, aunts, cousins, and grandparents for their love, sacrifices, and support throughout my academic journey. Their belief in me has been invaluable, and I am immensely grateful for having them in my life.

I am thankful to God and the Virgin Maria, for their watchfulness and guidance throughout my journey, and their presence, support, and unconditional strength, especially in moments of darkness and uncertainty.

I am also profoundly grateful to the furry and feathery angels who accompanied me throughout this research journey and provided me with unconditional love and comfort at the beginning and end of the day. Their presence in my life has been a source of joy and hope. Also, I extend my gratitude to my girls, Boni, Laika, and Woo Young Woo for their unwavering love, and loyalty, and for eagerly awaiting my return home.

This dissertation would not have been possible without every one of your contributions. From the bottom of my heart, thank you so much.

Resumo

Deep Learning tem mostrado excelentes resultados em tarefas de visão computacional, e a área de saúde não é exceção. *Deep Learning* pode auxiliar os dermatologistas no diagnóstico precoce de câncer de pele, o que pode salvar muitas vidas. No entanto, não há uma maneira direta de mapear o processo de tomada de decisão dos modelos DL. Para previsões de câncer de pele, não basta ter uma boa precisão; é necessário entender o comportamento do modelo para implementá-lo clinicamente e obter previsões confiáveis. Neste trabalho, identificamos desideratos para explicações em modelos de lesões de pele e apresentamos um estudo sobre como a *eXplainable Artificial Intelligence* está sendo usada atualmente para lesões de pele. Analisamos sete métodos (quatro baseados em atribuição de pixels e três baseados em conceitos de alto nível): Grad-CAM, Score-CAM, LIME, SHAP, ACE, ICE, CME para duas redes neurais profundas, Inception-v4 e ResNet-50, treinadas no *International Skin Imaging Collaboration Archive* (ISIC). Nossos resultados indicam que, embora essas técnicas mostrem efetivamente o que o modelo está procurando para fazer sua previsão, as explicações obtidas não são completas o suficiente para obter transparência nos modelos de lesão de pele.

Abstract

Deep Learning has shown outstanding results in computer vision tasks, and healthcare is no exception. Deep Learning (DL) can assist dermatologists in early skin cancer diagnosis, saving many lives. However, there is no straightforward way to map out the decision-making process of DL models. For skin cancer predictions, it is not enough to have good accuracy. Understanding the model's behavior is needed to implement it clinically and get reliable predictions. We identify desiderata for explanations in skin-lesion models and present a study about how eXplainable Artificial Intelligence (XAI) is currently used for skin lesions. We analyzed seven methods (four based on pixel-attribution and three high-level concepts): Grad-CAM, Score-CAM, LIME, SHAP, ACE, ICE, CME for two deep neural networks, Inception-v4 and ResNet-50, trained on the International Skin Imaging Collaboration Archive (ISIC). Our findings indicate that while these techniques effectively show what the model is looking to predict, the obtained explanations need to be completed more to get transparency into the skin-lesion models.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The rise of Artificial Intelligence (AI) has impacted several areas. It has expanded to developing automated (decision-making) systems without human intervention. Deep learning (DL) is a subfield of Machine learning (ML), which in turn is a field within AI [76].

DL models are the most known due to their superior performance. They consist of multiple layers connected through non-linear functions that automatically discover useful features representing the data's abstractions. Deep Neural Networks (DNNs) have shown promising results in different tasks and applications, such as image classification, object detection, semantic segmentation, visual question answering, natural language processing, speech recognition, audio processing, and many other well-known applications. However, there is a lingering black-box perception of DNNs, meaning that deep learning models can be assessed based on their final outputs without understanding how and why they make these decisions [68].

Decisions in critical and sensitive areas such as healthcare, legislation, law-making, housing, criminal justice, financial lending, employment, and driving cars were previously made only by human judgment. Now, they base their decision on the output of DNNs. However, how to entrust humans' life to a black-box system? To create trust and confidence in intelligent systems and to integrate them into our everyday lives, we must build "transparent" models that explain why they predict what they predict [80].

New approaches to explain how models make decisions have appeared to address this problem, creating a new research field named eXplainable Artificial Intelligence (XAI). Two approaches in the XAI literature help to understand DNNs: self-explanatory model and post-hoc explanations. A self-explanatory model can explain by itself its behavior to make a decision is interpretable by design. Conversely, post-hoc explanations are generated using external methods over a trained network. Explainability ensures impartiality in decision-making, facilitates robustness, and assures that only meaningful variables infer the output [5].

Melanoma is one of the most aggressive skin cancer and can be cured if detected early [105]. Deep Neural Networks (DNNs) have shown outstanding results for skin-lesion analysis [14, 16, 18, 24, 60, 66, 67, 98]. However, it is urgently needed to comprehend the model's behavior to implement these models clinically and save more human lives. In this Master's dissertation, we investigate explainability methods for skin lesion analysis.

1.1 Problem Statement

In medicine, common tasks are detection and recognition given a set of images. Conversely, although different explainability methods exist, they were initially developed for benchmark tasks that use ImageNet. Therefore, adapting existing methods for tasks involving medical images is challenging due to the significant difference in the data, such as data containing different types of animals and objects compared to pictures of skin parts. The homogeneity in medical images, such as skin lesion images, makes that obtained explanations different from what was expected.

Furthermore, the robustness of visual explanations has been assessed for generic domains (ImageNet [2]), and medical domains (brain MRI [31], radiology images [6]). The evaluation has been done by verifying the change in the explanation when the data label or the model parameters has been modified. Thus, it is possible to know whether the explainability method shows what is more important for the network to make a prediction or only identifying features, e.g., edges. However, the results in these three works are different. Some methods that proved robust for ImageNet were not for a medical domain. This makes us wonder what methods could be beneficial to explain DNNs in skin lesion analysis.

Moreover, recent studies [93,99] about XAI for the medical field agree with the need for more comparative studies of explainability methods. It is desired that these comparison studies show the inconsistencies between the different methods to have a realistic view of the current state of XAI in medicine and improve it. In that sense, we compare methods and assess their trustworthiness for skin lesion analysis.

1.2 Motivation and Challenges

According to the Global Cancer Observatory, in 2020, more than 19 million people worldwide were diagnosed with cancer, and the number of deaths surpasses 9 million people [91]. The new cases of melanoma, the most aggressive type of skin cancer, are more than 324 thousand people worldwide, and the number of deaths is superior to 57 thousand people [92]. The estimated number of new melanoma cases for 2025 exceeds 350 thousand people, while the estimated number of deaths because of melanoma is more than 63 thousand people.

Melanoma and other types of skin cancer can be cured if detected early [105]. According to the American Cancer Society [86], the 5-year survival rate after being diagnosed with melanoma in an advanced stage (cancer has spread to other parts of the body) is 30%, but if it is detected when it has started, the rate increases to 99%.

The ordinary flow to know if a lesion is a cancer starts with a clinical examination, and then, if there is a suspicion to be malignant, the patient is referred to a dermatologist who performs a visual inspection with a dermatoscope. Finally, if the dermoscopic examination of the image is cause for concern, a histological examination of the skin is conducted using a sample taken from the lesion (biopsy). The biopsy is an invasive and painful procedure that could lead to post-surgery complications. Furthermore, since it is time-consuming and expensive, it is challenging to access it. Thus, an automated system

for detecting skin cancer would be very beneficial for medical professionals and patients by determining urgency levels in triage rather than chronological order and supporting as assistance to diagnosis during routine visits or specialized consultations while also reducing the workload that has increased significantly since the COVID-19 pandemic.

Convolutional Neural Networks (CNNs) aimed to classify skin lesion images have been shown to perform reasonably, being superior to human raters(dermatologists and general practitioners) [20,38,93]. To put these models on deployment to really support specialists with the diagnosis and save more human lives, an understanding of the model's behavior is needed.

One of the most concerning issues when making predictions with DNNs is that the model relies its decision on spurious correlations. For instance, in our domain, the model could use non-lesion areas and external artifacts such as pen marks, bubble gels, color patches, and ruler marks rather than the presence of the attributes in the lesion. Therefore, many XAI studies for skin lesion analysis have focused on verifying that the model is looking at the lesion but still needs to solve the black-box issue proving the reasons and justification why models make particular decisions.

To get transparency and understanding on the inner workings of a DNN, the explanation should be informative enough to align with medical knowledge. In that sense, for the explanations to make sense and to build trust, we expect they contain dermoscopic attributes dermatologists use to diagnose and, of course, some coherent machine features. The annotation by experts for dermoscopic attributes is detail-focused and time-consuming, resulting in a very costly process. Currently, only one dataset contains around 2000 images annotated with their dermoscopic attributes, not only indicating their presence with yes/no labels but with masks that point out the exact location of the attribute in the lesion. With that in mind, we assess the obtained visual explanations from different methods compared to the presence of dermoscopic attributes in the lesion.

1.3 Objectives

The main objectives of this research are:

O1. To investigate how explainability methods have been used for skin lesion classification.

O2. To explore different explainability methods for skin lesion classification.

O3. To analyze and evaluate the obtained explanations.

1.4 Research Question

The main research question that this dissertation aims to answer is:

Q1. Considering that classifying skin lesion images is very different from common domains such as ImageNet due to input homogeneity, is it possible to apply current explainability methods to understand skin lesion models?

1.5 Contributions

We summarize our main contributions as follows:

C1. We provide a study about how explainability methods are being used for skin lesion analysis.

C2. We chose a desideratum for trustworthy explanations of DNN models that works with skin lesion analysis.

C3. We assess seven explainability methods in publicly available datasets; therefore results can be reproduced.

1.6 Outline

The remainder of this text is structured as follows. In Chapter 2, we describe the terminology and concepts around explainability and interpretability, types of explainability methods, and the methods used in this research. In Chapter 3, we review the literature for explainability methods used in skin lesion classification models. In Chapter 4, we describe the datasets, models, and methodology used to achieve our goals. Then, in Chapter 5, we present the experimental setup and the obtained explanations. Lastly, in Chapter 6, we summarize our findings and suggest future directions.

Chapter 2

Background

In this section, we will explore the differences between the terms interpretability and explainability, the importance of explainability in different areas, and the different types and forms of explainability methods. Also, we describe the methods we use in this dissertation.

2.1 Interpretability and Explainability

The terms *interpretability* and *explainability* have been used interchangeably in several works since they are closely related in their goal to explain the decision made by a model. However, it is essential to highlight the differences between both terms.

Molnar [64], Miller [62], and Biran and Cotton [13] define **interpretability** as "the degree to which a human can understand the cause of a decision"; while Kim et al. [48] define it as "the degree to which a human can consistently predict the model's result". Other definitions for interpretability are "the desirable quality or feature of an algorithm which provides enough expressive data to understand how the algorithm works" [30], and "the ability to explain or to provide the meaning in understandable terms to a human" [5].

While for interpretability, we were able to find many definitions; for **explainability**, there was only one that is "explainability are the details and reasons a model gives to make its functioning clear or easy to understand" [5]. Therefore, explainable AI is defined as "given an audience, an Explainable Artificial Intelligence produces details or reasons to make its functioning clear or easy to understand" [5].

The use of interpretability and explainability depends on the end-goal and the end-user. For our objectives, we will use these terms differently: Interpretability as an inherent feature in a model, i.e., the model explains by itself how it works; and Explainability as the process to turn a non-interpretable model into an explainable one [5]. In this sense, some models are self-interpretable, such as Logistic/Linear Regression, Decision Trees, K-Nearest Neighbors, Rule-based Learners, General Additive Models, and Bayesian Models. We will focus on explainability since we want to work on trained Deep Neural Networks (DNNs) that give good accuracy but are black-box.

2.2 Why is Explainability Necessary?

Successful DNNs are composed of several layers with many non-linear functions [90]. They compress the input features and then transform them into a weighted sum, followed by an activation function, and repeat this for many subsequent layers. Then, the decision is made based on the output of the DNN. This allows the network to learn different abstraction levels of the input and makes it difficult to trace and understand how the DNN makes the decision.

DNNs have been applied in several critical domains, e.g., medicine, healthcare, criminal justice, and financial lending. Therefore, there is a need to understand how they make their decisions to increase their confidence in them, to know whether a model is robust to adversarial attacks, to facilitate the detection of bias, and to assess whether a model is suitable for deployment.

Additionally, explainability is attached to legal and ethical concerns. Regulations such as the European General Data Protection Regulation (GDPR) [37] have opened the need for trustability, transparency, and fairness in Machine Learning.

Also, an explainability method allows for improving model performance, knowing the causes for a prediction, and using just the necessary features to have a good model.

2.3 Types of Explainability Methods

Among all the explainability methods, different categorizations can be seen in recent surveys [5, 77, 106, 115]. Here, we present different groupings based on Camburu [21] to provide a background for future work.

2.3.1 Post-hoc Explainability and Self-explanatory Models

This categorization is the most notable. It separates the explainability methods according to whether explanations are produced after the network training.

a. **Post-hoc** explanations are generated after the target model has already made the decisions; they explain trained models.

b. **Self-explanatory** models provide an explanation along with the model prediction; this could be more desirable than post-hoc methods. They train the predictor and the explanation method jointly, modifying the architecture or the optimization process of the given network to achieve that. Hendricks et al. [41] proposed early work on classification, providing text justification and image classification results. Self-explanatory models can sometimes result in higher or lower task performance, as was noticed by Camburu [21].

Since the models that have already performed well on skin lesion classification are not inherently interpretable, our first goal is to explain these trained networks and, from there, be able to get a self-explanatory model that does not lower the performance; but does increase trust. Therefore, we will focus on post-hoc explanations in our methodology.

2.3.2 Model-Agnostic and Model-Specific Methods

The difference between these categories relies on the model's knowledge to explain; this is mainly applicable to post-hoc methods.

a. **Model-Agnostic** methods only have access to the prediction of any input but not the architecture itself, so they are independent of the model's architecture and can be applied to any architecture. Some methods that fall in this category are LIME (Local Interpretable Model-agnostic Explanations) [71], Anchors [72], and Kernel SHAP (SHapley Additive exPlanations) [57]. These are also known as black-box explainers.

b. **Model-Specific** methods have access to the model's architecture and are mainly designed to work over a determined architecture. These are also known as white-box or model-dependent explainers. A few examples that belong to this category are Integrated Gradients [89], DeepLIFT (Deep Learning Important FeaTures) [84], and Grad-CAM (Gradient-weighted Class Activation Mapping) [80].

2.3.3 Local and Global Explainability Methods

The most common division among post-hoc explainability methods is local and global methods, i.e., whether the method tries to explain the behavior of the network as a whole or just a particular prediction.

a. **Local methods** explain the model's reasoning according to a specific input. The explanations will be analogous only for the specific input and those similar to it, but not for all the data points in the class. Thus, there are several different explanations for a group of data, and the user can conclude from these results for the whole group. This explanation is suitable when there is a need to unravel an individual prediction to an end-user.

b. **Global methods** explain the model's reasoning according to the class label or a determined set or neighborhood of inputs, i.e., the explanation is valid for all the data points in the class or set.

2.4 Forms of Explanations

Explainability methods can also be divided according to the explanation they produce. We will briefly describe the most used forms of explanation.

a. **Feature Importance Methods**: These methods are the most popular. They evaluate the importance of each feature of the inputs to a DNN. Some methods that measure the importance of each feature and are also model-agnostic are: LIME [71], Anchors [72], and Kernel SHAP [57]. For Computer Vision tasks, feature attribution methods are also known as visualization methods. These highlights, through a scientific visualization, characteristics or features of an input that strongly influence

the output of a DNN. For example, considering image classification, a good visual explanation should localize the image's target category (class-discriminative) and capture fine-grained details. This can be done through backpropagation, that is, getting the gradients through each layer and perturbation, altering or removing the input feature, and comparing the difference in network output between the original and altered one.

b. **Surrogate Explanations or Model Distillation**: These are methods where the knowledge in a trained DNN is extracted into a simpler representation (decision trees, finite state automata, graphs, or rule-based models). Distilled models can achieve reasonable performance, even being simpler because they have access to information from the trained DNN (more discriminatory input features or correlations in the output) and can use it for their training. They use the original data as input, developing a transparent model of how input features relate to the actions of the DNN. The resulting explanation can be seen as a hypothesis of why a DNN has assigned some class label to an input. Some related works on classification using CNNs were proposed by Harradon et al. [39] using autoencoders and by Zhang et al. [112] using graphs and trees [113].

c. **Concept-based Explanations**: These methods work similarly to feature attribution methods, but instead of assessing the importance of features, they assess concepts at a high level. These concepts can be defined by the user or can be learned from the input data. Most recent works based on this approach rely on a vector representative of the concept. Concept Activation Vector (CAV) is a normal to a hyperplane that separates examples without a concept and examples with a concept in the model's activations [46].

There are other forms of explanations, such as rule-based explanations, natural language explanations, prototype/example-based explanations, and counterfactual explanations, which are not presented since they are not necessary to understand the scope of this work. For more details, please refer to Arrieta et al. [5], Xie et al. [106], and Samek et al. [77].

In the following sections, we will explain the methods used in this dissertation. These are divided according to what they attribute the prediction: pixels or concepts.

2.5 Pixel Attribution Methods

Pixel Attribution methods produce a saliency map showing each pixel's importance for its classification label. There are several pixel attribution methods, and they can be classified into two types:

a. **Backpropagation-based Methods**: These methods distinguish the effect of input features on the final prediction based on some evaluation of gradient signals passed from output to input during network training. Saliency maps highlight the pixels in order of importance to the DNN prediction based on derivatives. Several works have

followed this line, some of these methods are Activation Maximization [32], Deconvolution [110, 111], Class Activation Mapping (CAM) [80, 116] and the subsequent methods Grad-CAM [80] and GradCAM++ [23] that generalize it, Layer-Wise Relevance Propagation (LRP) [7], Deep Learning Important FeaTures (DeepLIFT) [84], Integrated Gradients [89]. These methods generally have low quality and noise and lose information in the backpropagation process (vanishing gradient) due to Sigmoid and ReLU activations.

b. **Perturbation-based Methods**: These methods compare the difference in the prediction when the input features are changed. The most known methods are Occlusion by different types of perturbations [34] such as replacing the region with a constant value, adding noise to a region, and blurring a region; LIME [71], SHAP [57] and Smooth Masks [33].

In 2020, a combination of perturbation and Class Activation Mapping based methods was presented. Wang et al. [101] proposed a pixel attribution method, Score-CAM, based on CAM (Class Activation Mapping) and perturbation methods, giving a score to each activation map according to how each one will affect the prediction when applied as a mask to the input image.

Therefore, we chose methods representative of each type and Score-CAM using both approaches. These methods are described in the following.

2.5.1 Grad-CAM

Gradient-weighted Class Activation Mapping [80] (Grad-CAM) is a generalization of Class Activation Mapping [116] (CAM) and shows what parts of an image a CNN is looking at to give a prediction. CAM is a popular method that requires a specific architecture: after the last convolution layer, there must be a global average pooling (GAP) layer followed by a fully connected (FC) layer that will give the predictions. Therefore, to explain a particular class with CAM, the weights from the FC layer corresponding to the class are selected and multiplied with its corresponding activation map. However, to apply CAM to any deep network, the model must be changed and retrained to fulfill the architecture requirement; this can lead to a loss in performance. This problem is solved by Grad-CAM, which may be applied to a wide range of network architectures and tasks without the need for retraining. Grad-CAM uses the class-specific gradient information concerning the final convolutional layer of a model to produce a localization map of the most critical regions of the image for the prediction. Figure 2.1 illustrates Grad-CAM.

Grad-CAM works in the following way:

1. Feed the input image to the CNN;

2. Get the raw score y (before applying softmax activation) for the desired class c

3. Backpropagate the raw score for each pixel of each feature map activations A^k in the last convolutional layer: $\frac{\partial y^c}{\partial A^k}$;

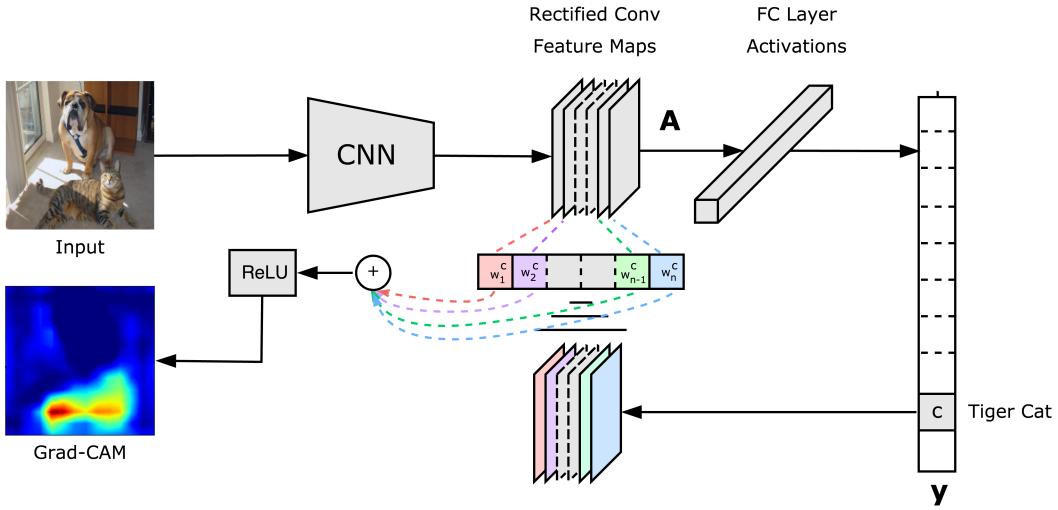


Figure 2.1: Outline of Grad-CAM: Given an image and the class 'Tiger cat', the image is fed to the CNN, and the value in the neuron associated with 'Tiger cat' is backpropagated to the activation maps in the convolutional layer. Then, a weighted combination of activation maps is performed with the average of gradients. The localization map (blue heatmap) is obtained after applying ReLU over the weighted combination. Figure reproduced from Selvaraju et al. [80].

4. To get the weight α_k^c, apply a global average pooling over the gradients (indexed by i and j) for each feature map activation A^k:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \tag{2.1}$$

5. Weight each activation map A^k with its corresponding α_k^c and average them;

6. Finally, the localization map produced by Grad-CAM is given by the equation below, where ReLU is applied to the averaged feature map to select features that impact positively in the class of interest:

$$L^c = \text{ReLU}(\sum_k \alpha_k^c A^k). \tag{2.2}$$

The main drawback of this method is the loss of information while performing the backpropagation process (vanishing gradients). This also happens in other gradient-based methods due to ReLU and Sigmoid activations.

2.5.2 Score-CAM

The idea behind Score-CAM [101] is similar to perturbation-based methods since modifications in the input will be performed, mand the difference in the prediction will be measured. The modifications in the input will be the activation maps in a specific layer of the network. These maps are applied as masks to the input image. Finally, a localization map (saliency map) is obtained. Figure 2.2 outlines Score-CAM.

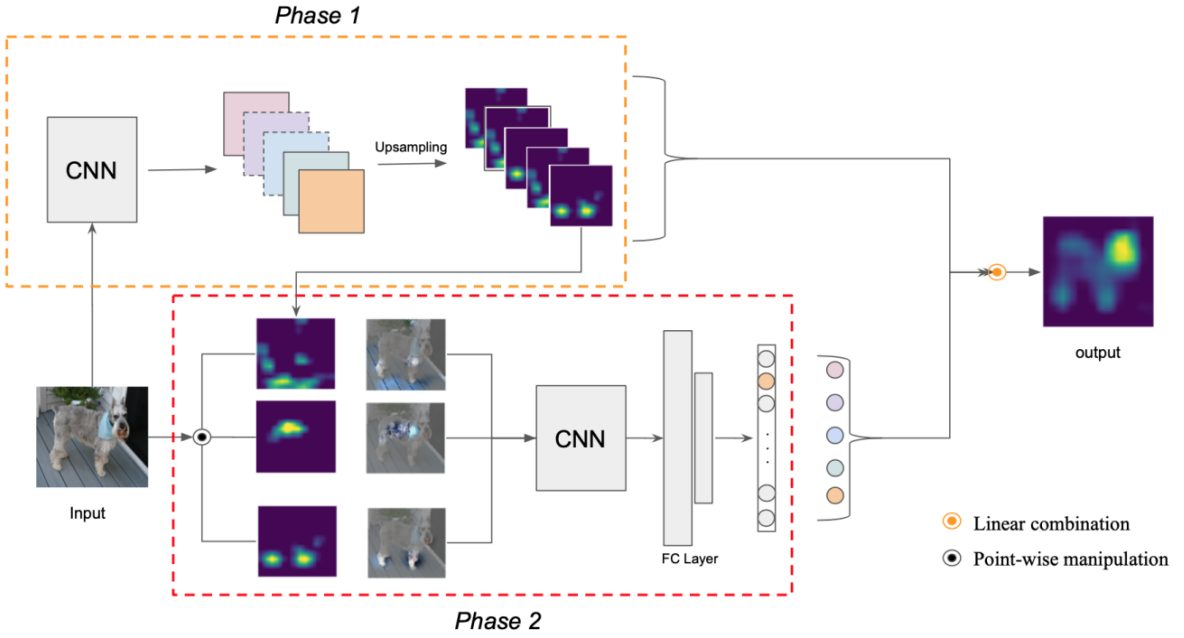


Figure 2.2: Score-CAM pipeline. Phase 1 gets the activation maps upsampled to the input size, and Phase 2 applies them as a mask to the input image. Figure reproduced from Wang et al. [101].

The steps to follow to get the localization map with Score-CAM for a prediction are as follows:

1. Pass an input image x to the CNN f;
2. Extract activation maps at a chosen layer l in the CNN;
3. Upsample activation maps to the original input size in the CNN;
4. Normalize values of each map A_l^k in the range of 0-1 (A_l^k denotes the activation map for the k-th channel in the l layer of the CNN), by applying:

$$\text{normalize}(A_l^k) = \frac{A_l^k - \min A_l^k}{\max A_l^k - \min A_l^k},$$

5. Apply the activation map as a mask to the input by multiplying it with the normalized activation map;
6. Pass masked inputs through the CNN and get the difference α_k^c in prediction with the input image. The differences work as a score for each map activation;
7. Apply softmax function to ensure all the differences sum up to 1;
8. Perform a linear weighted combination of all activation maps and their scores, and since we are interested in knowing the features that positively influence the prediction, a ReLU activation function is used.

$$L_{Score-CAM}^C \leftarrow \text{ReLU}(\sum_k \alpha_k^c A_l^k).$$

Score-CAM gets rid of the vanishing problem present in Backpropagation-based Methods. However, according to the chosen layer, the activation map can be of very small size, leading to a loss of information in the explanation. The obtained saliency map can be insufficient for some domains. Additionally, evaluation of the explanations can be tedious since it is a local method that produces a different explanation for each prediction.

2.5.3 LIME

Local Interpretable Model-Agnostic Explanations (LIME) [71] is a method that explains the outcome of black box machine learning models by using local surrogate models. Local surrogate models can be any interpretable model, e.g., Linear Regression or Decision Tree, that is trained to approximate the prediction of the black box model, for instance.

LIME creates a new dataset of perturbed samples and their corresponding predictions in the model. Then, an interpretable model is trained with this new dataset weighted according to the proximity to the original instance. Finally, the explanation comes from the interpretable model, e.g., in the case of linear regression, the explanation will be the interpretation of coefficients for the features.

Mathematically, given a black box model f,the explanation ξ for an instance x obtained with LIME [71] is expressed as:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_x) + \Omega(g), \tag{2.3}$$

where LIME tries to find the interpretable model g that minimizes the loss $\mathcal{L}(f, g, \pi_x)$, which measures the fidelity of g in approximating f for a locality defined by π_x while keeping the model complexity $\Omega(g)$ low. G is the family of potential interpretable models, and π_x is the proximity measure between a sample and x, and it is used to determine the size of the locality around the instance x that is considered for the explanation.

The steps to get the explanation with LIME for a prediction from a model trained with images are listed below and illustrated in Figure 2.3:

1. Segment the image in different areas(superpixels);
2. Create a new dataset with perturbed images, i.e., some superpixels are turned off (pixels change to zero);
3. Get the black box prediction for each perturbed image;
4. Calculate the weights for the new samples by measuring the proximity to the original image. The greater the closeness between a disturbing picture and the original image, the greater the weight and relevance of the perturbed image;
5. Train an interpretable model with sample weights and the perturbed samples where each superpixel is considered to be a binary feature. Weights are used when computing the loss and will encourage the model to be more precise in predicting samples closer to the original image;

6. Map more important features from the interpretable model to their corresponding superpixels.

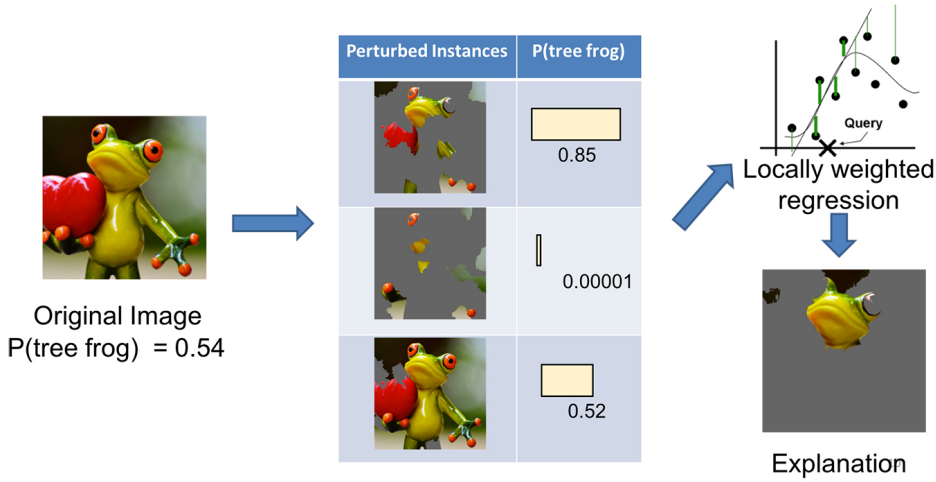


Figure 2.3: Explaining the prediction "tree frog" with LIME for classification task using ImageNet. First, generate perturbed instances and find their prediction on the black box model. Second, pass them to a weighted regression and gets the explanation from it. Figure reproduced from Ribeiro et al. [70].

As LIME is model-agnostic, it can explain any black box model. It works with tabular data, images, and text. Local Surrogate Models can use other features than those used in the original model but must be obtained from the data, e.g., apply transformations in the images instead of using superpixels. Some of the significant LIME downsides are: it is vulnerable to adversarial attacks, which means it can be used to create deliberately deceptive explanations [85] and the fact that explanation for the same prediction changes each time LIME is executed [64]. Therefore, LIME is very unreliable and unstable method.

2.5.4 SHAP

SHapley Additive exPlanations (SHAP) [57] is an agnostic method that can be applied to any ML model. To get the feature attributions, it uses Shapley Values [81], a method from cooperative game theory that tries to find a fair distribution of the payout between all players in a group.

The Shapley value for a player i calculates the weighted average of the marginal contributions of player x to the payout. For this, it gets all subsets of players that do not contain player i, then it computes the marginal contribution of i as the effect (difference) on the payout when player i is added to all subsets, and finally aggregates all contributions.

Mathematically, the Shapley value φ_i of a player i is defined by:

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{(N - |S| - 1)!|S|!}{N!} [v(S \cup \{i\}) - v(S)], \tag{2.4}$$

where S is a subset of features(players), $v(S)$ is the prediction (payoff) for the coalition or subset S, N is the total number of features, and $N \setminus \{i\}$ is all the possible subsets

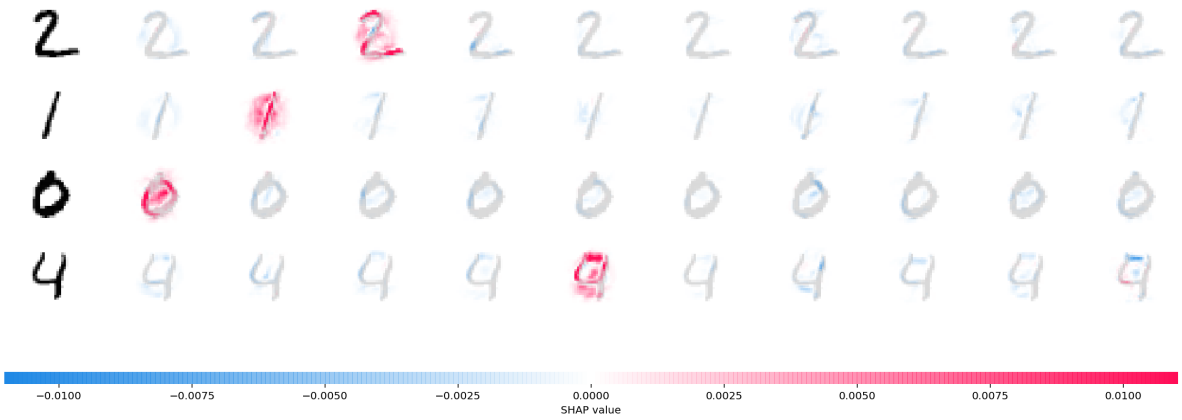


Figure 2.4: Examples of explanations for MNIST predictions using SHAP: Deep Explainer. The pixels in red are the ones that contributed to increasing the prediction result. Figure reproduced from Lundberg [56].

not containing feature i. The marginal contribution of i in the subset S is given by $v(S \cup \{i\}) - v(S)$.

Shapley Values can be applied to get feature attributions in Machine Learning by associating players with features and the payoff with the prediction. Shapley Values in their original form find the average marginal contribution to the end result, however for ML, it will usually find the average contribution for a background group that works as a 'baseline' for the explanation, e.g., Why was the tumor malignant compared to all benign tumors?

In Machine Learning, the models cannot just exclude one of its features to make a prediction. Thus, to see how the output changes when a feature does not participate in the prediction, SHAP represents the missing features as unknown values and mimics the unknown scenario by averaging all possible values for the feature. Furthermore, computing Shapley values following Equation 2.4 is very expensive since its complexity is $2^{\text{number of features}}$. To deal with these problems, SHAP uses a sample of all possible coalitions to estimate the Shapley Values. Figure 2.4 depicts some explanations obtained with SHAP.

SHAP [57] includes different extensions (Tree Explainer, Gradient Explainer, Linear Explainer, Kernel Explainer, and Deep Explainer) according to how to approximate the Shapley Values in different ML architectures. Here, we describe two of them since they will be used in this research:

- **Kernel Explainer:** This approach uses a surrogate model to estimate Shapley Values. Here, the explanation g is represented as:

$$g(x') = \phi + \sum_{j=1}^{M} \phi_j z'_j,$$

where $z \in 0, 1^M$ is a coalition vector indicating with one the presence of a feature and with zero its absence, M is the maximum coalition size, and ϕ_j is the feature attribution for a feature j. The steps to find the contribution of each feature in a

model f for an instance x are:

- Sample coalitions or subsets, and represent them in a coalition vector z_k;
- Get prediction for each z_k by first converting it to the original feature space;
- Compute the weight for each coalition z_k using $\frac{(M - |z_k| - 1)!|z_k|!}{M!}$, small and large coalitions are given a large weight;
- Fit the weighted linear model;
- Return Shapley values ϕ_k, the coefficients from the linear model.

- **Deep Explainer:** This implementation is a faster algorithm to compute SHAP values for deep learning models based on connections between SHAP and the DeepLIFT [84] algorithm. For more details, please refer to Shrikumar et al. [84].

SHAP is well known for having a solid theoretical foundation in game theory [64]. It can produce local and global explanations, fairly distribute the prediction among features, and get contrastive explanations. However, although efforts to compute Shapley Values faster have been made, it still needs to be a faster algorithm. Another drawback is that it needs access to all training data to deal with features' exclusion. It also ignores feature dependence giving weight to points very unlikely to exist.

2.6 Concept-based Methods

Concept-based methods started with Bau and Zhou [12], which used visual concepts found in the data through segmentation and evaluated their alignment in the hidden units. Next, we describe the concept-based methods we use.

2.6.1 TCAV

When presenting an image with its features highlighted according to the weight in the decisions to humans, what they do is, try to find patterns between explanations, qualitatively identify parts (high-level concepts), e.g., the tail in a cat, the strips in a zebra, the ears of dogs.

The idea behind TCAV (Testing with Concept Activation Vectors) [46] is to map a vector space, where the features and neural activations of the model are, to another different vector space with vectors corresponding to concepts that humans are related to. In this work, the user provides the concepts by giving a set of samples.

In perturbation-based methods, features were used as a form of perturbation in the input, and then, the network's response was checked. In this way, TCAV is a global perturbation method since it produces explanations for each class.

Figure 2.5 outlines TCAV. To calculate the CAV, the authors first selected a layer in our neural network, considering that higher layers encode more abstract information since their receptive field is bigger while lower layers are more specific. Suppose we are working on layer l, the examples and counterexamples of concept C are fed to the neural

network till layer l. In this way, we obtain the activations at layer l. Next, a binary linear classifier is trained to distinguish between the two sets of activations that belongs to the examples and the counterexamples. The classifier is the CAV for the given concept.

In saliency maps, we measured the sensitivity of a class for a pixel with derivatives. TCAV computes the sensitivity concerning the direction of a concept, i.e., a CAV, on layer l for a data point x. The conceptual sensitivity of class k for the concept C is given by the directional derivative $S_{C,k,l}$:

$$S_{C,k,l} = \nabla h_{l,k}(f_l(x)) \cdot v_C^l, \tag{2.5}$$

where $f_l(x)$ are the activations of layer l for a data point x, $h_k(x)$ is the prediction for x for class k, and v_C^l is a CAV of concept C in layer l. Lastly, we calculate the TCAV score by measuring the sensitivity for all input data of class k and calculating the ratio of input data with a positive directional derivative:

$$TCAVQ_{C,k,l} = \frac{|x \in X_k : S_{C,k,l}(x) > 0|}{|X_k|}. \tag{2.6}$$

The main caveat of this approach is the dependency on human-defined labeled concepts that can lead to bias.

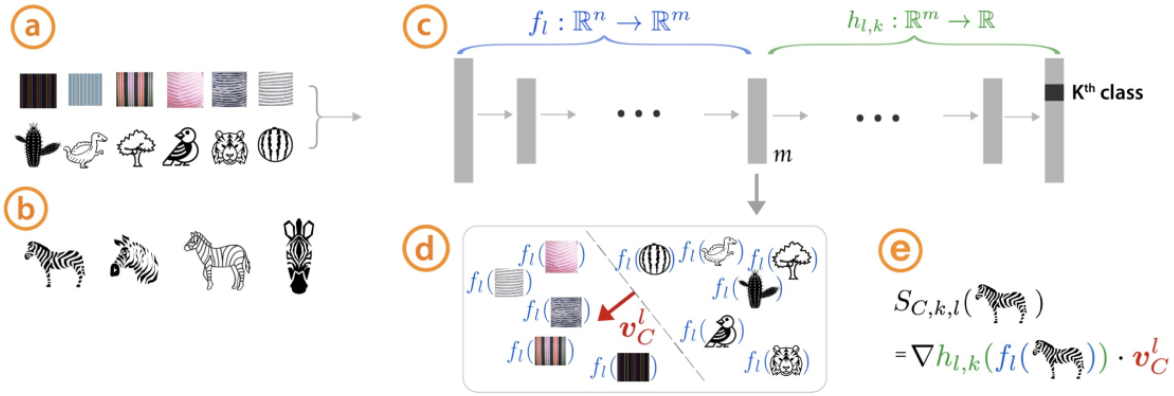


Figure 2.5: Outline of Testing with Concept Activation Vectors (TCAV) method for concept-based explanations. Example using given sets of samples for the concept 'striped' ((a) first-row), random samples ((a) second-row), training-data samples for class 'zebras' (b), and a pre-trained network (c). Concept Activation Vectors (CAVs) are learned by training a linear classifier with the activations produced by the concept's samples and the random samples in a determined layer of the model (d), then, in order to quantify the sensitivity to the concept for a class, it is used a directional derivative (e). Figure reproduced from Kim et al. [46].

2.6.2 ACE

As stated, TCAV needs annotated concept samples. Thus, in 2019, Ghorbani et al. [36] developed the Automatic Concept-based Explanations (ACE) method to overcome this dependence. The focus of this work is to get the concept samples automatically since

getting human-labeled concept samples is costly. For this, it uses superpixels on input data to form concept samples. Figure 2.6 outlines ACE.

The algorithm for ACE is as follows:

1. Segment given class images to superpixels with multiple resolutions;
2. Upsample each superpixel to the model's input size;
3. Feed the superpixels to the model and get its activation on a chosen layer;
4. Measure the Euclidean distance between segments' activations;
5. Cluster similar segments;
6. Remove outlier segments by removing those with low similarity to the cluster center;
7. Each cluster represents a concept; therefore, the importance of each concept is measured using TCAV.

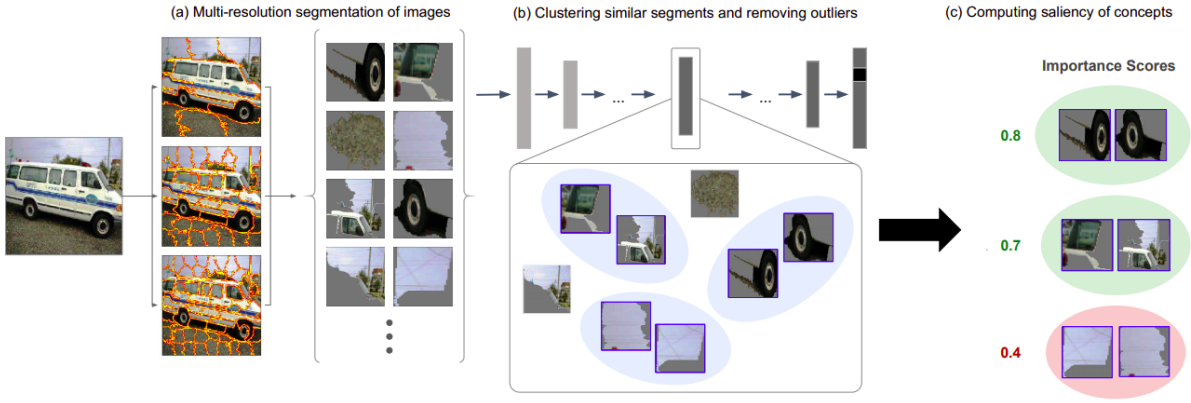


Figure 2.6: ACE algorithm. (a) A set of images from the same class is given. Each image is segmented with multiple resolutions resulting in a pool of segments from the same class. (b) The activation space of one bottleneck layer of a CNN classifier is used as a similarity space. After resizing each segment to the model's standard input size, similar segments are clustered in the activation space, and outliers are removed to increase the coherency of clusters. (d) For each concept, its TCAV importance score is computed given its examples segments. Figure reproduced from Ghorbani et al. [36].

The main limitations of ACE are that concept information can be lost when deleting outliers and that since concepts are retrieved from a sample of images, there is no certainty that these concepts are enough to explain the model accurately.

2.6.3 ICE

Invertible Concept-based Explanations (ICE) [115] proposes an unsupervised method to improve ACE to have concept explanations (global and local). It gets better results in interpretability and fidelity, which means it can get a prediction from the concepts

using a linear approximation (invertible explanations). For that, it uses non-negative matrix factorization (NMF) in the activation maps of the CNN feature extractor to find the CAVs (Concept Activation vectors) and gets its importance using the weights in the linear approximation. Figure 2.6 outlines ICE.

Non-negative matrix factorization (NMF) is used to reduce the number of features and to find its representative vector (CAVs). Given a CNN, let A of shape $n \times h \times w \times c$ be the activation or feature maps produced by the filters of size $w \times h$ and c channels from a target layer l for all the n images in the training set. For ICE, A is assumed to be non-negative since most CNNs use the ReLU activation. Feature maps A can be flattened to $V \in R^{(n \times h \times w) \times c}$, NMF reduces the channel dimensions of V from c to c'. Therefore, $V = SP + U$, where $S \in R^{(n \times h \times w) \times c'}$ indicates the scores(how much they are related to P) and $P \in R^{c' \times c}$ is a vector basis indicating components that repeatedly appears across all n data points and serve as the essential building blocks from which we can approximately rebuild all of the original data points. NMF is implemented as an optimization problem that minimizes the residual error U.

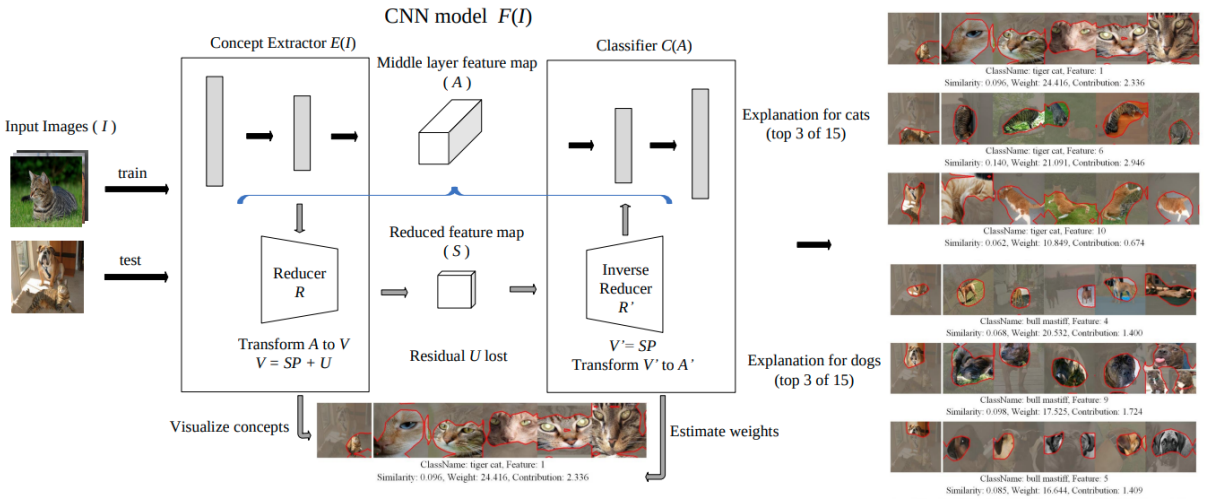


Figure 2.7: ICE algorithm. First, a middle layer is chosen to divide the CNN model into a feature extractor and classifier. Feature maps at the chosen layer are reduced using NMF. The decomposed feature maps are used to visualize concepts, and weights from the classifier part are used along with the reconstructed feature maps to find the importance of concepts. Figure reproduced from Zhang et al. [115].

In summary, the steps to get explanations for ICE are as follows:

1. Separate the CNN into feature extractor and classifier to choose a layer. The authors use the last layer since it contains high-level features;

2. Flatten feature or activation maps A at layer l as $V \in R^{(n \times h \times w) \times c}$;

3. Reduce channel dimension in V using non-negative matrix factorization $V = SP + U$, P will be the set of Non-negative Concept Activation Vectors (NCAVs);

4. Assuming that after the last convolutional layer l there is a GAP layer and a dense

layer, prediction for class k can be retrieved from concepts as:

$$\begin{aligned} C_k(A_l) &= \text{GAP}(A_l)W + b \\ &= \text{GAP}(S)PW + \text{GAP}(U)W + b, \end{aligned}$$

where $W \in R^c$ are the weights from the dense layer for target class k and b is the bias;

5. Global feature importance for NCAV P and class k will be PW;

6. To find local explanations for a new image, apply NMF on the known P to compute a new S with the new feature maps. S can be considered the degree of similarity to NCAVs in P. Thus, to find the contribution of each concept in an image prediction, the similarity score is multiplied by the concept weight;

7. Visualize the features using prototypes (images with the highest similarity scores to the concept) and highlight the area representing the concept by employing the decomposed feature map as a heatmap for a single CAV with a threshold.

The ICE authors showed that their approach gets better results in fidelity when getting the vectors representing the concepts than other methods, such as clustering. Also, ICE is more interpretable than clustering and PCA (Principal Component Analysis). However, some explanations could be incorrect since approximate models are being used, and fidelity is not always 100%. Additionally, access to all training data is necessary.

2.6.4 CME

Concept-based Model Extraction (CME) [44] is a framework that allows the analysis of already pre-trained DNNs by explaining how the model uses concept information when making predictions. For this, they use model distillation, i.e., information (internal representations) from the trained model is used to predict concepts, and it uses them with a transparent model (Logistic Regression or Decision Trees) to mimic the actions of the DNN. Figure 2.8 depicts CME.

CME explores if a DNN is concept-decomposable. The authors [44] gives the following definition: "A DNN f is concept-decomposable if it can be well-approximated by a composition of functions p and q, such that $f(x) = q(p(x))$, where the function $p : \mathcal{X} \to \mathcal{C}$ is an *input-to-concept* function mapping data points from their input representation $x \in \mathcal{X}$ to their concept representation $c \in \mathcal{C}$, the function $q : \mathcal{C} \to \mathcal{Y}$ is a *concept-to-output* function mapping data-points in their concept representation $\mathcal{C}$ to output space $\mathcal{Y}$". Thus, CME tries to approximate f with an extracted model $\hat{f}(\mathbf{x}) = \hat{q}(\hat{p}(\mathbf{x}))$. We can identify two main functions in CME:

- **Concept Predictor** ($\hat{p}$): This function uses the activations from CNN middle layers to predict concepts. To do this, the authors assume at least a small set of training data have labels for k concepts. Concept predictor $\hat{p}$ is composed of a set of functions that predict the presence of each concept, as outlined below:

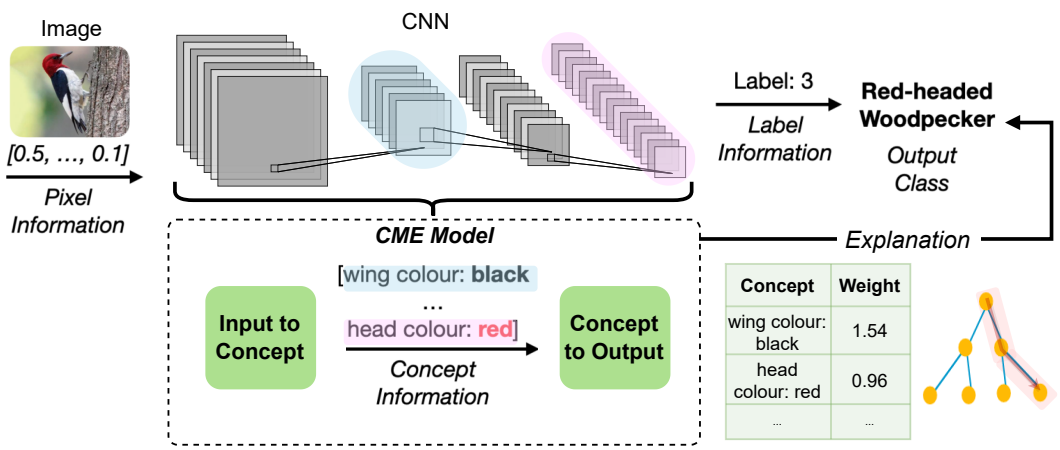


Figure 2.8: CME outline. The input image is transformed into concept information from CNN's internal representation. Each concept is predicted from the layer which extracts it with the minimum error. Then, an interpretable model is trained from predicted concepts to get the original predicted label. Thus, the explanations could be the weights on a linear model or a decision tree path. Figure adapted from Kazhdan et al. [44].

1. Given a layer l in a CNN model, get the layer's representation (activations) of the training data $\mathbf{h}$;
2. Use Semi-Supervised Multi-Task Learning (SSMTL) to train predictors for each concept from $\mathbf{h}$. Each concept is handled as a separate and independent task;
3. Repeat steps 1 and 2 for all model layers getting a set of functions $G = \{g_i^l | l \in \{1 \dots L\} \wedge i \in \{1 \dots k\}\}$ and select the function with the lowest error rate for each concept.

- **Label Predictor** ($\hat{q}$): CME trains an interpretable model ($\hat{q}$) with the predicted concepts generated by ($\hat{p}$) to get the output labels. These models could be Logistic Regression or Decision Trees.

CME allows intervention on the concepts the model learned, improving task performance. Also, CME is designed to be data efficient, i.e., it only needs some training data to have concept labels and shows good results on tasks related to synthetic datasets. However, for a real-life scenario such as bird species prediction task using the CUB dataset [100], the fidelity and task performance were low, indicating that the CUB model was *non-concept-decomposable*, which implies the model could not be explained with the desired concepts (the model relies on other non-concept information). Additionally, other work [45] showed that the performance to predict concepts depends on how much the concept affects the end task label.

Chapter 3

Related Work

In this chapter, we review current works that use explainability methods for skin-lesion diagnosis. Hauser et al. [40] reviewed how explainable AI was being used for dermoscopic data. They selected 37 studies from 2017 since the first deep learning approach that got results at the dermatologist level was published at the beginning of 2017 till September 2021.

We drew inspiration from Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) [63] for our analysis. The articles were selected from different manual searches on Google Scholar. We decided to limit our search between 1st January 2020 and 3rd June 2022 since previous works are already reviewed by Hauser et al. [40] and to have a manageable amount of papers to analyze. We excluded surveys and reviews. All included articles perform classification tasks, and to get a fair comparison, all use DNN's excluding other ML approaches. The queries were performed as follows:

- A query with the terms 'interpretability skin lesion dermoscopic clinic explanation -preprint -audio -video' gave 256 results. We excluded works that did not perform a classification task nor use any explainability method, remaining 14 studies.
- A query with the terms 'interpretability skin lesion dermoscopic clinic explanation xai' gave 60 results. After analyzing each work, only two new studies remained.
- A query with the terms 'interpretability skin lesion dermoscopic clinic explanation' resulted in 675 works. After inspecting each result till page 8, three studies remained.

Finally, we added nine works from 2020 to 2021 that did not appear on the previous searches but did appear in the systematic review. Also, we included three additional works to the set of selected works, totaling 31 works. Tables 3.1 to 3.1 reviews 31 works that apply XAI methods, divided according to the form of explanation, example-based: prototype (provides images similar to the input as explanation), example-based: counterfactuals (provides examples on how the input has to change its prediction), pixel attribution (provides heatmaps with pixels highlighted according to its importance for the prediction), and concept-based (provides human-understandable concepts). Also, the table details the code availability for reproducibility, the usage of XAI, and how the explanations are evaluated.

Nearly 66% (21/31) of the reviewed articles used pixel attribution methods, i.e., CAM, Grad-CAM, LIME, SHAP, and Attention. Around 50% (14/31) of the studies used XAI superficially: 12/15 as a sanity check, i.e., to show that the classifier focuses on the lesion or parts of it without any further analysis or evaluation, 2/15 as a sanity check (pipeline) that included the XAI method in the model, so it can be self-explainable (gives the prediction and its explanation) but without any further analysis too.

From all the selected articles, about 9% (3/31) presented a new XAI method and showed how it could be used in skin lesion models, and close to 19% (6/31) of the studies improved previous XAI methods, e.g., improving the resolution of the saliency map [82], turning an ante-hoc XAI method into a post-hoc method [109]. Nearly 22% (7/31) of the articles presented a detailed analysis and evaluation of their results using XAI methods for skin lesions models. Only 25% (8/31) of the works presented quantitative metrics to evaluate the explanation. The most common quantitative way to asses pixel-attribution methods is to compare segmentation masks versus the saliency map, i.e., check how many important pixels are in the lesion.

Only one [88] of the reviewed studies performed a comparison between three explainability methods: LIME, Grad-CAM, and SHAP applied on a Vanilla CNN trained with the HAM10000 dataset. Results were evaluated qualitatively using the criteria: fidelity, consistency, sensitivity, and clinical relevance.

Our work explores different explainability methods focusing on concept-based explanations since they provide interpretable information. In the last part of Table 3.1, we can find four works that use concepts as explanations. We can divide these works into two groups: those that use CAVs and those that do not.

Three works belong [54, 55, 109] to the first group. The first one [54] uses TCAV for a model that classifies Melanoma and Seborrheic keratosis. It trains the concept classifiers on PH2 [59] and Derm7pt [43] datasets for 11 concepts in each dataset. PH2 consists of only 200 images with concept annotations and their corresponding mask. Derm7pt consists of 823 dermoscopic images (Melanoma and Nevi) with annotated concepts without masks. Similarly, the second work ExAID [55] extends [54] by combining learned CAVs with Concept Localization Maps to get not only the most important concepts but their localization in images. The third work [109] adapts Concept Bottleneck Models to be used post-hoc. For this, they used learned CAVs with Derm7pt and trained an explainer that uses data projections over the CAVs. There is an implicit assumption that the concepts the explainer will use are easier to learn than the concept or class it tries to explain. This means that having predicted concepts with lower performance than the class can put in doubt the correctness of the explanation [69]. All of the works mentioned above present the majority of their concepts to be more challenging to learn the label itself, making us wonder if the obtained explanations are reliable. This could result from learning concepts with small datasets such as PH2 and Derm7pt.

In the second group, there is only a work [87] that customizes LIME, a perturbation method, to be used with the ABCD rule by modifying the images along with diagnostic characteristics. Their experiments used only border and color, showing how the model had learned these two concepts when making a prediction. However, extending this approach to other diagnostic characteristics and dermoscopic attributes is challenging.

Therefore, the literature reviewed shows concept-based explanations for the skin lesion model, but these still need to be more reliable and interpretable.

Table 3.1: Overview of reviewed works that apply XAI for skin lesions classification (Continued).

	Ref_{Year}	Code	Usage of XAI	Method	Target Model	Task	Datasets	Explanation Evaluation	Metric
Example-based Explanation: Prototype	$[95]_{2020}$		Analysis on human interaction	CBIR, Grad-CAM	ResNet34	7 cls[a]	HAM10000	Qualitative, Quantitative	User engagement, Pixel activation (background vs lesion area)
	$[4]_{2020}$		Technical analysis	CBIR	ResNet50	8 cls[b]	ISIC2019	Qualitative	–
	$[10]_{2021}$	✓	Improvement of XAI method	CBIR	DenseNet121, ResNet101, VGG16	7 cls[a]	ISIC2018	Qualitative	–
	$[11]_{2021}$		Improvement of XAI method	CBIR	ResNet50	7 cls[a]	ISIC2018, ISIC2019	Qualitative	–
Example-based: Counterfactuals	$[61]_{2021}$		new XAI method	ABELE	ResNet50	9 cls[c]	ISIC2019	Qualitative	–
	$[35]_{2022}$	✓	New XAI method	DISSECT	DenseNet, AlexNet	2 cls: melanoma vs. benign lesions	SynthDerm	Qualitative, Quantitative	Importance, Realism, Distinctness, Stability and Substitutability
Pixel Attribution	$[58]_{2020}$	✓	Improvement of XAI method	Gradient, Integrated Gradients	ResNet18	3cls: MEL, NV, BKL	HAM10000	Qualitative	–
	$[9]_{2020}$	✓	Sanity check (pipeline)	Visual Attention	DenseNet161, ResNet50, VGG16, + LSTM	hierarchical classification	ISIC2017, ISIC2018	Qualitative	–
	$[73]_{2020}$	✓	Improvement of XAI method	Grad-CAM	VGG16	2 cls: malignant, benign	ISIC Archive	Qualitative	–
	$[52]_{2020}$		sanity check	CAM	AlexNet, DenseNet161, PNASNet5, ResNet50, SENet154, VGG19	7 cls[a]	ISIC2018	Qualitative	–

[a] AK/IEC, BCC, BKL, DF, MEL, NV, VASC
[b] AK/IEC, BCC, BKL, DF, MEL, NV, VASC, SCC
[c] AK/IEC, BCC, BKL, DF, MEL, NV, VASC, SCC, UNK

Table 3.1: Overview of reviewed works that apply XAI for skin lesions classification (Continued).

	Ref_{Year}	Code	Usage of XAI	Method	Target Model	Task	Datasets	Explanation Evaluation	Metric
Pixel Attribution	[106]$_{2020}$	✓	sanity check (pipeline)	CAM	MB-DCNN	3 cls: MEL, NV, SK	ISIC2017 Task 2, PH2, ISIC Archive (1320 images)	Qualitative	–
	[117]$_{2020}$	✓	sanity check	Grad-CAM	VGG16	2 cls: MEL, NV	ISIC2016	Qualitative	–
	[42]$_{2020}$		sanity check	t-SNE, Grad-CAM, LIME	Inception-ResNetV2	2 cls: BCC, SK	Chinese skin dataset	Qualitative	–
	[50]$_{2020}$		sanity check	Grad-CAM, Guided Grad-CAM	ResNet101, ResNeXt, SEResNet, SEResNeXt and ensemble	7 cls[a]	HAM10000	Qualitative	–
	[88]$_{2020}$		elaborate analysis (comparative study)	LIME, Grad-CAM and Kernel SHAP	vanilla CNN	7 cls[a]	HAM10000	Qualitative	Fidelity, Consistency, Sensitivity and clinical relevance
	[103]$_{2020}$		sanity check	t-SNE + feature activation map	modified DenseNet121 and MobileNetv1, ensemble of both	2 cls: MEL and benign	ISIC2016	Qualitative	–
	[26]$_{2021}$		sanity check	CAM, Attention maps	custom 5-layer CNN (with attention)	7 cls[a]	HAM10000	Qualitative, Quantitative	Dice score: segmentation vs attention maps
	[65]$_{2021}$		technical analysis	Grad-CAM	ResNet50, VGG16	2 tasks: 8 cls[b]on ISIC19 and 2 cls: MEL, NEV on ISIC18	ISIC2019, ISIC2018 Task 2	Qualitative, Quantitative	IoU Coverage (Cvrg.) also known as Jaccard index

[a] AK/IEC, BCC, BKL, DF, MEL, NV, VASC
[b] AK/IEC, BCC, BKL, DF, MEL, NV, VASC, SCC
[c] AK/IEC, BCC, BKL, DF, MEL, NV, VASC, SCC, UNK

Table 3.1: Overview of reviewed works that apply XAI for skin lesions classification (Continued).

	Ref_{Year}	Code	Usage of XAI	Method	Target Model	Task	Datasets	Explanation Evaluation	Metric
Pixel Attribution	[82]$_{2021}$		Improvement of XAI method	HR-CAM	ResNet50, VGG19	8 cls[b]	BCN-20000, HAM10000, MSK, ISIC2017	Qualitative, Quantitative	sensitivity and specificity over the masks
	[83]$_{2021}$		Sanity check	SHAP	Ensemble of EfficientNetB0, DenseNet121, Xception	2 cls: MEL, NV	ISIC2020	Qualitative	–
	[97]$_{2021}$		Sanity check	Integrated Gradients	EfficientNetB3–B6 , Ensemble	2 cls: malignant, benign	ISIC2019, ISIC2020	Qualitative	–
	[107]$_{2021}$		Analysis on robustness	Attention	PMI2019AttnMel, InceptionV3	2 cls: MEL, others	ISIC Archive	Qualitative	–
	[102]$_{2021}$		Sanity check (pipeline)	SHAP (metadata, features), Grad-CAM (images)	EfficientNetB5	7 cls[a]	HAM10000	Qualitative	–
	[3]$_{2022}$		Sanity check	SHAP	MobileNet, Xception, ResNet50, ResNet50V2, and DenseNet121	2 cls: malignant, benign	ISIC2018	Qualitative	–
	[104]$_{2022}$		Sanity check	Grad-CAM	CNN + attention	2 classses: MEL, Non-MEL, 7 cls[a], 8 cls[b]	ISIC2017, ISIC2018, ISIC2019	Qualitative	–
	[51]$_{2022}$		Sanity check	CAM	MelaNet (custom CNN)	2 cls: MEL, Non-MEL	ISIC2017, PH2, MED-NODE	Qualitative	–
	[19]$_{2022}$	✓	Analysis on model vs human reasoning	LIME	ResNet50	2 cls: malignant, benign	ISIC2016	Qualitative, Quantitative	IoU Cvrg. (Jaccard), Ground Truth Cvrg, and Saliency Cvrg.

[a] AK/IEC, BCC, BKL, DF, MEL, NV, VASC
[b] AK/IEC, BCC, BKL, DF, MEL, NV, VASC, SCC
[c] AK/IEC, BCC, BKL, DF, MEL, NV, VASC, SCC, UNK

Table 3.1: Overview of reviewed works that apply XAI for skin lesions classification.

	Ref_{Year}	Code	Usage of XAI	Method	Target Model	Task	Datasets	Explanation Evaluation	Metric
Concept Attribution	$[54]_{2020}$		Analysis on concepts	TCAV	Inceptionv4 (RECOD)	3 cls: SK, MEL, NV	PH2 and Derm7pt for concepts, ISIC2017 for classification	Qualitative	–
	$[87]_{2021}$		Analysis on ABCD-rule	LIME	MobileNet	2 cls: MEL, NV	HAM10000	Qualitative	–
	$[109]_{2022}$		Improvement of XAI method	Post-Hoc CBM	Inception	2 cls: malignant, benign	HAM10000, SIIM-ISIC concepts: Derm7pt, Fitzpatrick17k	Qualitative, Quantitative	Metashift
	$[55]_{2022}$		New XAI method	ExAID: TCAV, CLM and textual explanations	SEResNeXt	2 cls: MEL, NV	Derm7pt, PH2, ISIC2016, ISIC2017, ISIC 019, SkinL2	Qualitative	–

[a] AK/IEC, BCC, BKL, DF, MEL, NV, VASC
[b] AK/IEC, BCC, BKL, DF, MEL, NV, VASC, SCC
[c] AK/IEC, BCC, BKL, DF, MEL, NV, VASC, SCC, UNK

Chapter 4

Methodology

In this chapter, we present the methodology used to explore and verify the usefulness of different explanatory methods for skin lesion analysis. We described the methodology in Section 4.1. Then, in Section 4.2, we detail the datasets. In Section 4.3, we specify the models used to explain. Finally, in Section 4.4, we explain how we select the explanatory methods and their details.

4.1 Pipeline

Figure 4.1 depicts the pipeline used in this research.

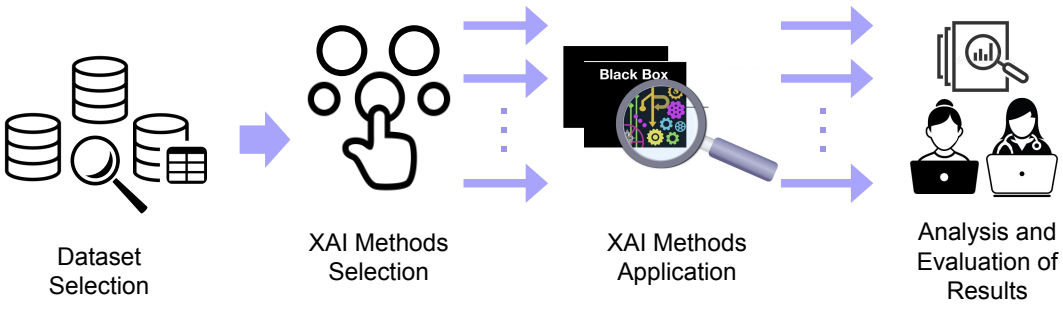


Figure 4.1: Proposed Pipeline.

1. **Dataset Selection** In order to evaluate visual explainability methods, we need a dataset with dermoscopic attributes and their localization in the image. Thus, we could compare the obtained results to the images with their dermoscopic attributes. Therefore, we chose ISIC 2018 Task 2 because of the dermoscopic attributes annotation (*pigment network*, *negative network*, *streaks*, *milia-like cysts*, and *dots/globules*, see Figure 4.2). To train the networks to explain, we chose HAM10000 since it is one of the most frequent datasets used in related works. Also, it does not present an intersection with ISIC 2018 Task 2 dataset. Thus, we can have a realistic idea of how models will behave in real-word since they will predict images that do not belong to the training dataset domain, i.e., images acquired in different light settings and the presence of new artifacts.

2. **Explainability Methods Selection** We can choose the methods from a huge pool. To choose pixel-attribution methods, we selected the most well-known robust methods (based on backpropagation and perturbation) that passed the sanity check, i.e., if the prediction changes, does the explanation change? For concept-based methods, we first verified the availability of code, then, whether the method could be applied to our domain and available datasets. For example, some methods require a large number of samples to learn a concept [45], therefore they were not used.

3. **XAI Method Application** After selecting explainability methods, we reproduced each work to validate the author's results. Then, we adapted the provided code to work with the architecture of our networks and performed several experiments to find the best hyperparameters for each method to explain the melanoma networks using the selected datasets. We described the experimental setup in Section 5.1. Also, since we are working on a medical domain (skin lesion analysis) and images are very similar, we need high-resolution results to find specific characteristics and patterns. Therefore, the visual explainability methods were adapted to provide explanations with good visual quality.

4. **Analysis and Evaluation of Results:** We assessed the quality of obtained explanations. Different properties help to judge how good explanations are. We can identify three desired properties that explanations should accomplish to make skin lesion models understandable:

    - Fidelity: It is the degree that indicates how well the explanation approximates the behavior of the black box model [64, 74]. The explanation should reflect what a model really does [78], and the obtained importance scores should be true, e.g., if the explanation says that altering a region in the input will change the model's prediction, it does. This property is also known as faithfulness [79] and accuracy [22]. Depending on the XAI method type, it is commonly evaluated quantitatively by inserting and removing the most relevant features according to the explanation and monitoring the change in the prediction. For surrogate models, the fidelity is given by the approximation quality of predictions.
    - Meaningfulness: "How well do users understand the explanations?" [64] "Does the explanation make sense?" [79]. Thus, the explanation must be expressive enough so that the users can associate it with some meaning, ensuring comprehension of the classifier's decision strategy. This property is also referred to as comprehensibility [64, 74] and understandability [22, 78].
    - Effectiveness: Given the explanation and the input, the user can simulate the outcome of the model [79]. Therefore, the explanation should be detailed enough to justify the decision, allowing the user to generate hypotheses to test. This property is also referred to as sufficiency [78].

4.2 Datasets

This section describes the datasets used in this Master's dissertation. Even though different lesion classes are provided in the datasets, we only use melanoma and benign lesions images since our purpose is to explain a melanoma classifier.

4.2.1 ISIC 2018 Challenge – Task 2

The International Skin Imaging Collaboration (ISIC) releases a challenge with different tasks and datasets each year. The ISIC 2018 Challenge [28] consisted of three tasks: (1) lesion segmentation, (2) lesion attribute detection, and (3) disease classification. It contains 2594 images with 12,970 ground-truth segmentation masks for dermoscopic attributes, 519 melanoma images, and 2075 benign images. We used this dataset to evaluate the explainability methods. Figure 4.2 shows an image per class with the localization of their respective dermoscopic attributes. Figure 4.3 shows some examples per class in this dataset.

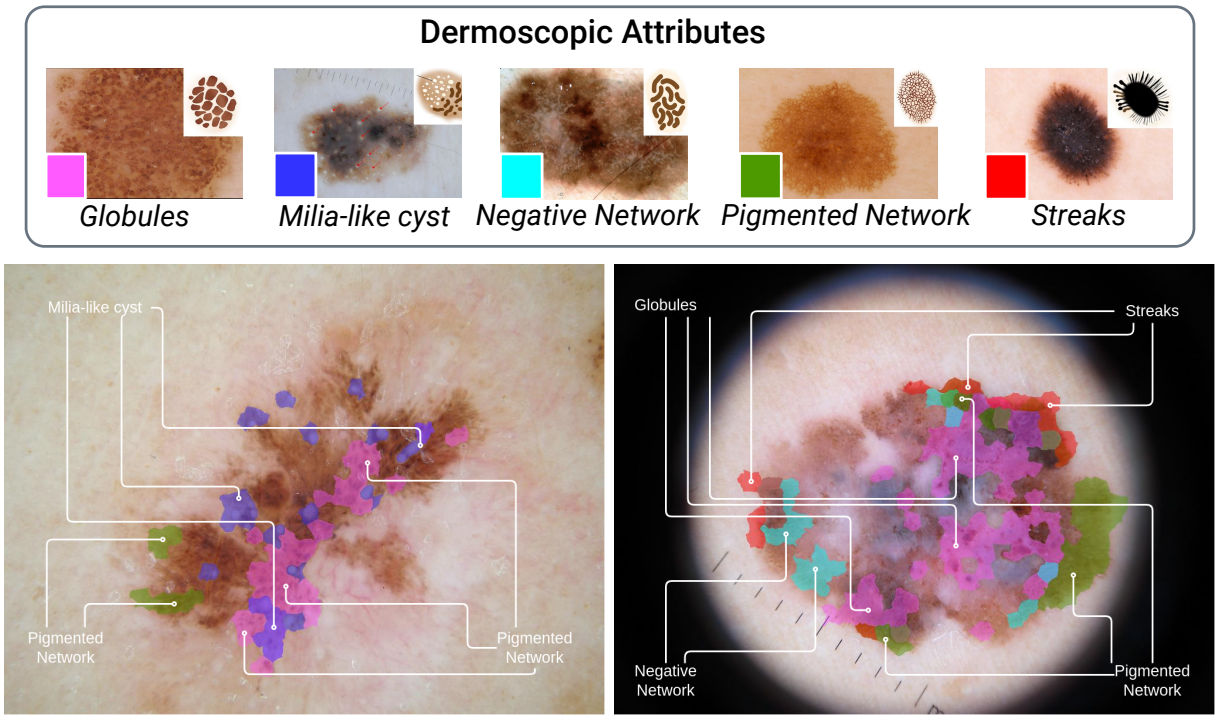
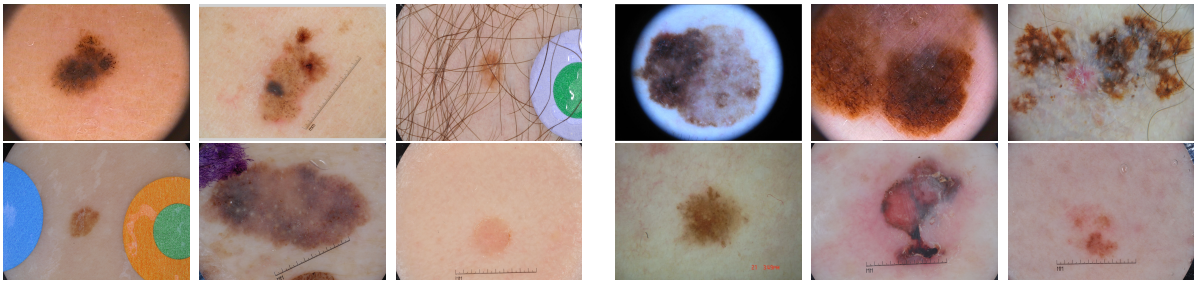


Figure 4.2: Presence of dermoscopic attributes in the benign lesion (left) and melanoma (right).

4.2.2 HAM10000

We trained our models with images from the HAM10000 dataset [94]. These images are also provided in the ISIC 2018 Challenge – Task 3: Disease Classification. We removed carcinomas images (i.e., malignant) and split the remaining images into a training set 70% (6,513 images), validation set 10% (930 images), and test set 20% (1,860 images).



(a) Benign lesions (b) Melanoma lesions

Figure 4.3: ISIC 2018 Task 2 dataset sample images.

HAM10000 dataset presents various images for the same lesion at different magnifications and angles (see Figure 4.4), so we make sure not to put these images in different sets, for instance, all three images of the lesion in Figure 4.4 belong to training set and no other set. Figure 4.5 shows some examples per class.

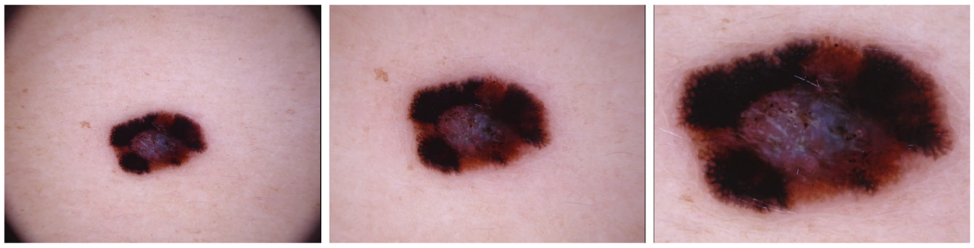
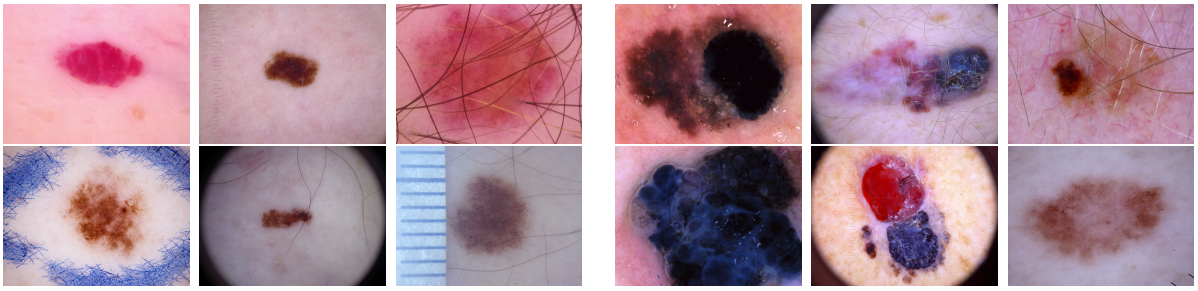


Figure 4.4: Example of different perspectives (magnifications and angles) for the same image lesion in HAM10000. Figure reproduced from Tschandl et al. [96].



(a) Benign lesions (b) Melanoma lesions

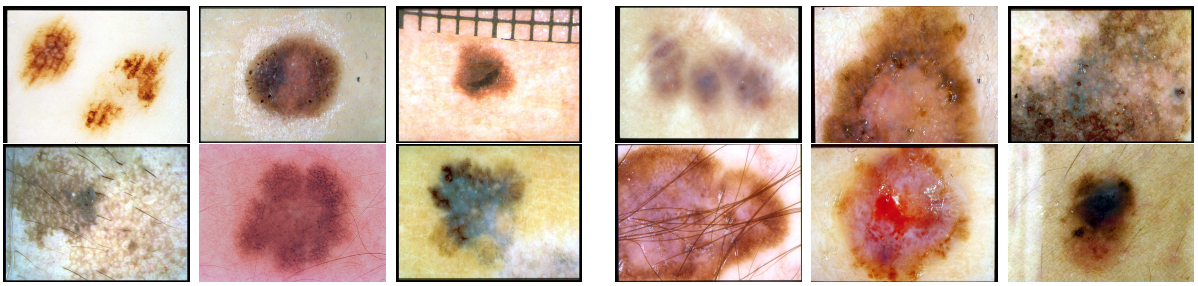
Figure 4.5: HAM10000 dataset sample images.

4.2.3 Derm7pt

We used 870 dermoscopy color images from the 7-point criteria evaluation dataset (Derm-7pt) [43], divided into training, validation, and test sets. We use the criteria as concepts. After concept preprocessing (i.e., turning concepts into binary, excluding concepts with only one value and low frequency), we ended up with 25 concepts. Table 4.1 lists the extracted concepts and corresponding values.

Table 4.1: Concepts and values used from the Derm7pt dataset.

Name	Values
Pigment network	absent, atypical, typical
Dots and globules	absent, irregular, regular
Streaks	irregular, regular
Blue whitish veil	present
Pigmentation	absent, diffuse irregular, diffuse regular, localized irregular
Regression structures	absent, blue areas, combinations, white areas
Vascular structures	absent, arborizing, comma, dotted, hairpin, linear-irregular, within regression



(a) Benign lesions (b) Melanoma lesions

Figure 4.6: Derm7pt dataset sample images.

4.3 Models

The target models we chose to explain are a skin-lesion classification Inception-v4 because of its well-known performance in skin-lesion analysis [98], and a ResNet-50 because of its popularity for skin-lesion classification as shown in Table 3.1. Both models are pre-trained on ImageNet in which we performed a fine-tuning with stochastic gradient descent with a momentum of 0.9, weight decay of 0.001, and learning rate of 0.001, reduced using a plateau scheduler that monitors validation loss with a patience of 10 epochs, reduction factor of 10, and minimum learning rate of 10^{-5}. The models were trained for a maximum of 100 epochs with early-stopping with a patience of 22 epochs monitored on validation loss. We used a batch size of 32, shuffling the data before each epoch. Input images were resized to 299×299 pixels for Inception-v4 and to 224×224 for ResNet-50, and z-normalized with ImageNet's training set mean $[0.485, 0.456, 0.406]$ and standard deviation $[0.229, 0.224, 0.225]$.

We applied basic data augmentation on train and validation with horizontal and vertical flips, rotations in the range of $45°$ to $-45°$ degrees, resized crops containing $75-100\%$ of the original image, and random changes in the brightness in the range of -40% to $+40\%$. We also used test-time augmentation, where we got 49 augmented versions of each test image, and to obtain the final prediction, we performed an average of 50 predictions.

The Inception-v4 model obtained an average of $89.96\% \pm 0.52$ Area Under the Receiver Operating Characteristic Curve (ROC AUC) over 6 runs, and the ResNet-50 got $90.37\% \pm 0.82$ ROC AUC over 6 runs also.

The code we used to train the models is in the repository deconstructing-bias-skin-lesion[1].

4.4 Explainability Methods

Table 4.2 summarizes the chosen methods. We chose gradient-based, perturbation-based, and CAM-based methods for pixel attribution methods. For works based on concepts, we can find that the literature works with different types of concepts annotation: visual (images indicating what parts of the image are related to each concept or specific samples representing a concept) and textual (images and tabular data of concepts per image). Some works [25, 49] build models interpretable by design using the textual concept annotations in an end-to-end training, and to get post-hoc explanations, we found different options with available code as TCAV [46], ACE [36], Completeness-aware Concept-Based Explanations [108], ICE [114], and CME [44]. We chose not to use TCAV since it requires samples to learn each concept, and there is no dataset with sufficient images and masks to learn dermoscopic attributes. Also, unfortunately, we could not run the code for Completeness-aware Concept-Based Explanations [108]. ACE and ICE try to find visual concepts automatically, and CME works with textual concept annotation.

[1] `https://github.com/alceubissoto/deconstructing-bias-skin-lesion`

Table 4.2: Overview of selected explainability methods.

	Ref_{Year}	Method	Type of Exp.	Target Model	Task	Datasets	Explanation Evaluation	Metric
Pixel Attribution	$[71]_{2017}$	LIME	Local	Inception, SVM	Image Classification, Text Topic Classification	ImageNet, Atheism and Christianity news	Qualitative	–
	$[57]_{2017}$	SHAP	Local	CNN	Image Classification	MNIST digit dataset	Qualitative	–
	$[80]_{2017}$	GradCAM	Local	VGG-16, VGG-16 + LSTM, CNN + LSTM	Image Classification, Image Captioning, Visual Question Answering	ILSVRC 2015	Qualitative, Quantitative	Localization error
	$[101]_{2020}$	ScoreCAM	Local	VGG16	Image Classification	ILSVRC 2012	Qualitative, Quantitative	Avg Drop, Avg Increase, Insertion and Deletion Scores, Sanity Check
Concepts attribution	$[36]_{2019}$	ACE	Global	Inception-v3	Image Classification	ILSVRC 2012	Qualitative, Quantitative	Smallest sufficient concepts (SSC), Smallest destroying concepts (SDC)
	$[114]_{2019}$	ICE	Local, Global	ResNet-50, Inception-v3	Image Classification	ILSVRC 2012, CUB	Qualitative, Quantitative	Fidelity, Understanding, Satisfaction, Sufficiency, Completeness
	$[44]_{2020}$	CME	Local, Global	custom CNN, Inception-v3	Image Classification	dSprites, CUB	Qualitative, Quantitative	Fidelity

Chapter 5

Experimental Results

In this chapter, we present the experimental results obtained with different explanatory methods applied to models trained to classify skin lesions. We split the chosen methods according to what they attribute to the prediction:

a. Pixel-attribution Methods (Section 5.2)

- Grad-CAM (Gradient-weighted Class Activation Mapping) [80], a method based on activation maps and gradients.
- Score-CAM (Score Class Activation Mapping) [101], a method based on activation maps and perturbations.
- LIME (Local Interpretable Model-Agnostic Explanations) [71], a method based on superpixels and perturbations.
- SHAP (SHapley Additive exPlanations) [57], a robust method based on game theory.

b. Concept-attribution Methods (Section 5.3)

- ACE (Automatic Concept-based Explanations) [36], a method that finds concepts automatically using clustering and superpixels.
- ICE (Invertible Concept-based Explanations) [115], a method that finds concepts using non-negative matrix factorization.
- CME (Concept-based Model Extraction) [44], a method that mimics the model behavior using an interpretable model with concept information.

All experiments were conducted on a single NVIDIA Quadro RTX 8000 GPU, with 48 GB of GDDR6 (Graphics Double Data Rate 6) synchronous dynamic RAM.

5.1 Experimental Setup

Since some methods work on PyTorch and others on TensorFlow, we converted the PyTorch models to TensorFlow using pytorch2keras[1]. Thus, we used the same learned

[1] `https://github.com/gmalivenko/pytorch2keras`

weights and got the same prediction. In the following, we describe the hyperparameters used for each method.

- We chose the last convolutional layer of both models for Grad-CAM, Score-CAM, ACE, and ICE. The implementation used for Grad-CAM and Score-CAM is available on Github[2].
- For LIME, we used the library implemented by the authors[3], we used a Ridge Regression linear model, a cosine distance function, and an exponential kernel. The saliency feature set is created using the top 5 features (superpixels created with QuickShif) that positively impact the model's prediction.
- For SHAP, we used Kernel Shap from the author's implementation[4] with the same superpixels used on LIME.
- For ACE, we followed the authors [36], we selected a random set of 50 images in the melanoma class, and to represent the random concept in the statistical significance test, we chose 50 images of the whole ISIC 2018 dataset; likewise, we chose 50 random images for each of the 50 random sets. We performed a SLIC (Simple Linear Iterative Clustering) superpixel segmentation [1] with 15, 50, and 80 segments to get the concepts' patches. These segments are completed with a gray value of 117.5 and passed through the networks to get their representation on the last layer. We used k-Means with $k = 25$ to cluster the representations and find the concepts. We removed clusters with few elements. For the TCAV score, the *p-value* is 0.05, so concepts with *p-value* greater than 0.05 have not passed the statistical significance test. Since the original code[5] from the authors was on TensorFlow version 1, an upgraded version on Github[6] was used and modified to be run with the selected models.
- For ICE, we used the implementation on Github[7] provided by the authors [114], we chose randomly 78 images per class, NMF is trained with a limit of 200 iterations, 16 components and 64 as batch size.
- For CME, we used the code implementation on Githhub[8] we chose the last 5 layers from which we learn concepts. Logistic regression for input to a concept is trained with a maximum of 200 iterations. We used Minimal Cost-Complexity Pruning for the Decision Tree with an alpha value of 0.00333.

5.2 Pixel-attribution Methods

Figures 5.1 and 5.2 show two examples (each one presents a skin-lesion image, an image with dermoscopic attributes, and a superimposed image) from the pixel-attribution

[2] https://github.com/yiskw713/ScoreCAM
[3] https://github.com/marcotcr/lime
[4] https://github.com/slundberg/shap
[5] https://github.com/amiratag/ACE/
[6] https://github.com/monz/ACE/tree/tensorflow-2-upgrade
[7] https://github.com/zhangrh93/InvertibleCE
[8] https://github.com/dmitrykazhdan/CME

methods according to the correctness of the prediction: (a) true positive, the classifier correctly predicts the melanoma class, (b) true negative, the classifier correctly predicts the non-melanoma class, (c) false positive, the classifier predicts benign as melanomas, and (d) false negative, the classifier predicts melanomas as benign.

We split Figures 5.1 and 5.2 into two rows. In the first row, most of the methods highlight the skin lesion, while in the second row the method focuses on spurious correlations, i.e., visible artifacts [15, 17] introduced during the image acquisition process, e.g., patches, gel bubbles (Figure 5.2c, second row, 4th, 5th and 6th columns), ruler marks (Figure 5.1a, second row, 4th, 5th and 6th columns, Figure 5.1d, second row, 3rd and 4th columns, and Figure 5.2d, second row, 3rd column), skin hair (Figure 5.1b, second row, 3rd, 4th, 5th and 6th columns, Figure 5.2b, second row, 3rd, 4th and 5th columns, Figure 5.2c, second row, 3rd, 4th, 5th and 6th columns, and Figure 5.2d, second row, 3rd, 4th and 5th columns). Here, the model bases its decision on surrounding parts of the skin lesion and even on other artifacts, such as a ruler. These model behaviors are certainly not desirable in neural networks that will support whether a person is healthy or not.

These results show what pixels are most important for the prediction. LIME is more specific than Grad-CAM and Score-CAM, selecting the essential superpixels for the predictions. However, when Grad-CAM and Score-CAM highlighted only the lesion, LIME also chose parts of skin surrounding the lesion (Figure 5.1c, first row, 5th column and Figure 5.2a, first row, 5th column, Figure 5.2b, first row, 5th column, and Figure 5.2d, first row, 5th column) and even only parts of skin without lesion (Figure 5.1b, first row, 5th column, and Figure 5.1d, first row, 5th column). On the other hand, SHAP shows in green the superpixels that contributed positively to the prediction and in red the superpixels that contributed negatively. Thus, for example, in Figure 5.1b, second row, we have a benign image predicted correctly where we could observe SHAP highlighted most superpixels surrounding the lesion in green, indicating that these pixels push the prediction to be benign, note how some of these areas that contain skin hair are presented as an explanation for SHAP, similar to Grad-CAM and Score-CAM. Here, there is a spurious correlation, as mentioned before.

These explanations are insufficient because they only say, “the classifier predicts the melanoma class because there is a skin lesion”. However, it does not provide any further information to support the prediction.

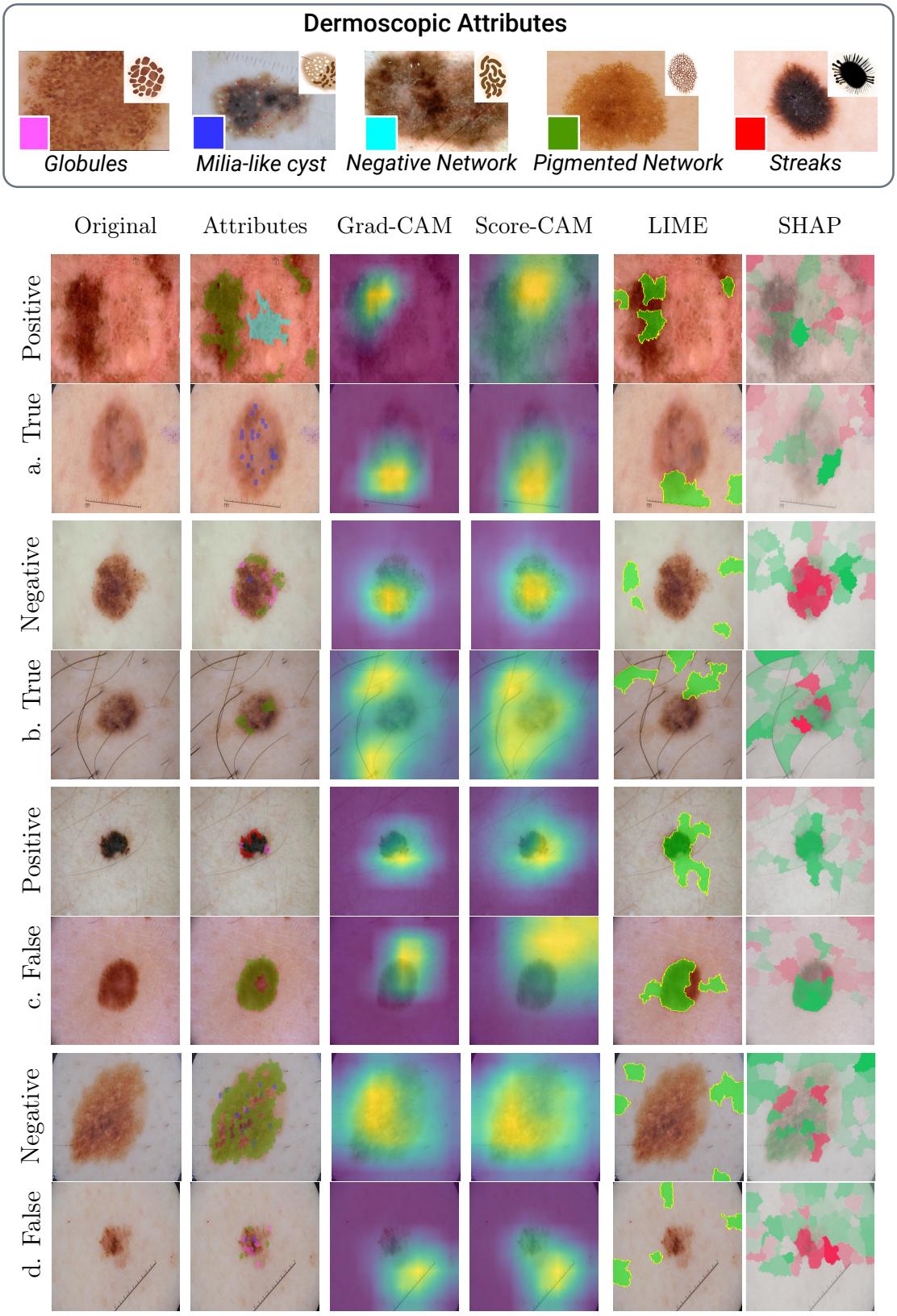


Figure 5.1: Saliency results for Inception-v4. Yellow colors in maps represent the parts of the input images that were more relevant for the prediction. For LIME, green parts represent the image that contributed positively to the prediction. For SHAP, pixels in green are those which contributed positively, and pixels in red are those which contributed negatively.

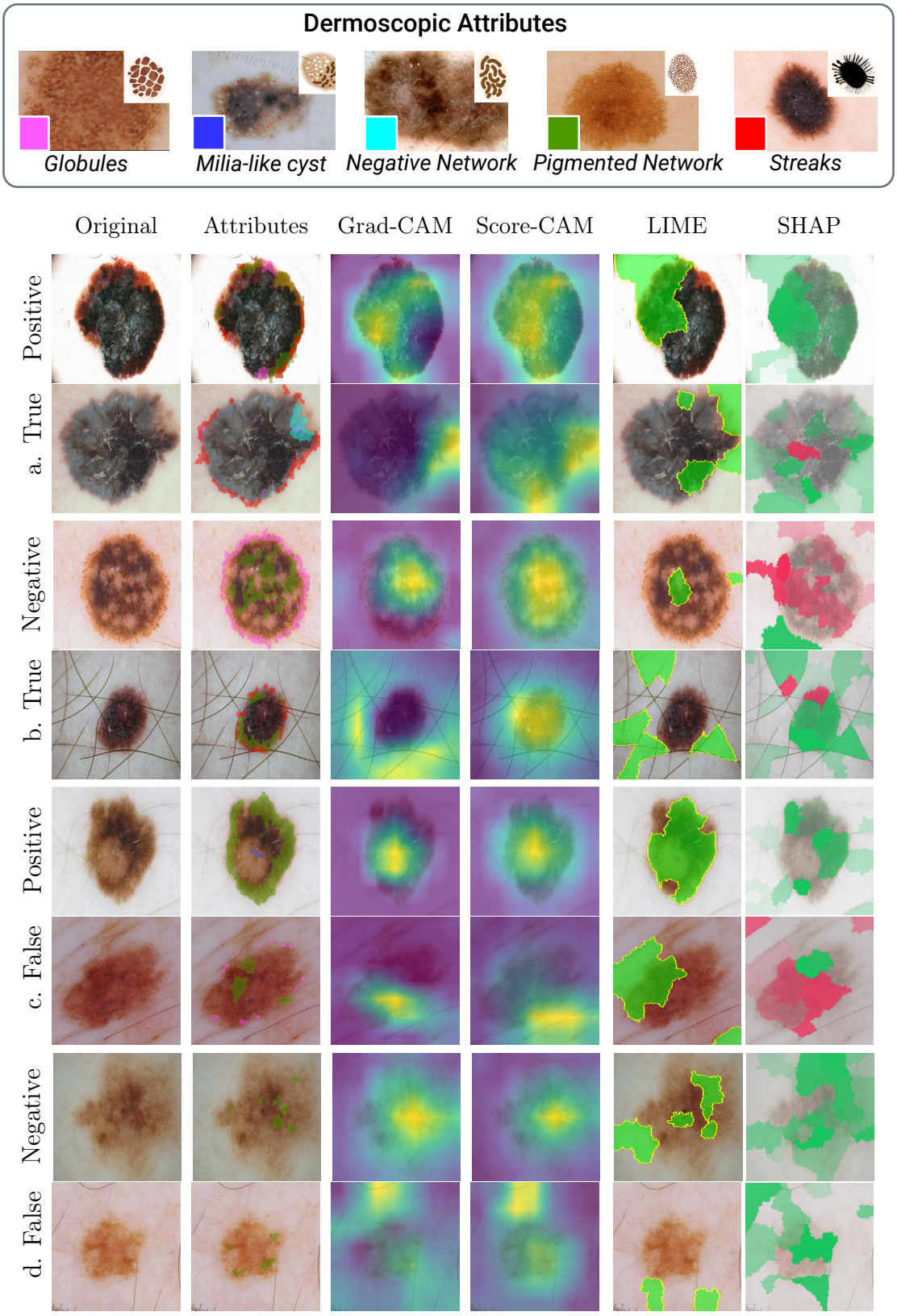


Figure 5.2: Saliency results for ResNet-50. Yellow colors in maps represent the parts of the input images that were more relevant for the prediction. For LIME, green parts represent the image that contributed positively to the prediction. For SHAP, pixels in green are those which contributed positively, and pixels in red are those which contributed negatively.

5.3 Concept-attribution Methods

In this section, we present results for explanations based on concepts: ACE and ICE find the concepts and their localization in an unsupervised setting, and CME uses textual concepts without localizing them. These three methods provide global explanations, but only ICE provides local ones. We recall that the networks we explain have a considerable performance, above 89% AUC-ROC on the HAM10000 test set. However, when tested on the datasets we use to contrast the explanations against dermoscopic attributes, ISIC 2018 Challenge – Task 2 and Derm7pt, the performance is lower (around 80% AUC-ROC). This explains how these networks will behave in a real-world setting. Therefore, for ICE and ACE, which uses representative images of a class to find the concepts, we chose those images predicted correctly by both networks.

5.3.1 ACE

Figure 5.3 and Figure 5.4 show ACE results for melanoma images predicted correctly (true positive), i.e., why the model is predicting all this set of images as melanoma according to ACE. We show six samples from the four most important concepts. Each sample shows the superpixel, the superpixel location in the image, and the dermoscopic attributes.

We found it hard to interpret the obtained concepts. At first sight, the most salient concept from the Inception-v4 model seems to be related to the pigmented network attribute. However, this attribute is also present in other concepts. This behavior is repeated on results from ResNet-50. This could be because this attribute is present on most of the images in the evaluation dataset.

While the first three most salient concept samples focus on lesion parts, in the fourth most salient concept, we can note samples from the skin that is around the lesion on both models: Inception-v4 (Figure 5.3, 4th Most Salient Concept, third column) and ResNet-50 (Figure 5.3, 4th Most Salient Concept, first row). This is confusing since more concept samples, even those into the lesion, are close to the lesion border, which could mean this concept is related to the lesion border.

The main drawback of ACE is that the concepts and scores can vary according to the chosen images for the target class, random concepts for the significance test, and the initialized weights to get the CAVs. Here, we present the results that got more concepts into the skin lesion.

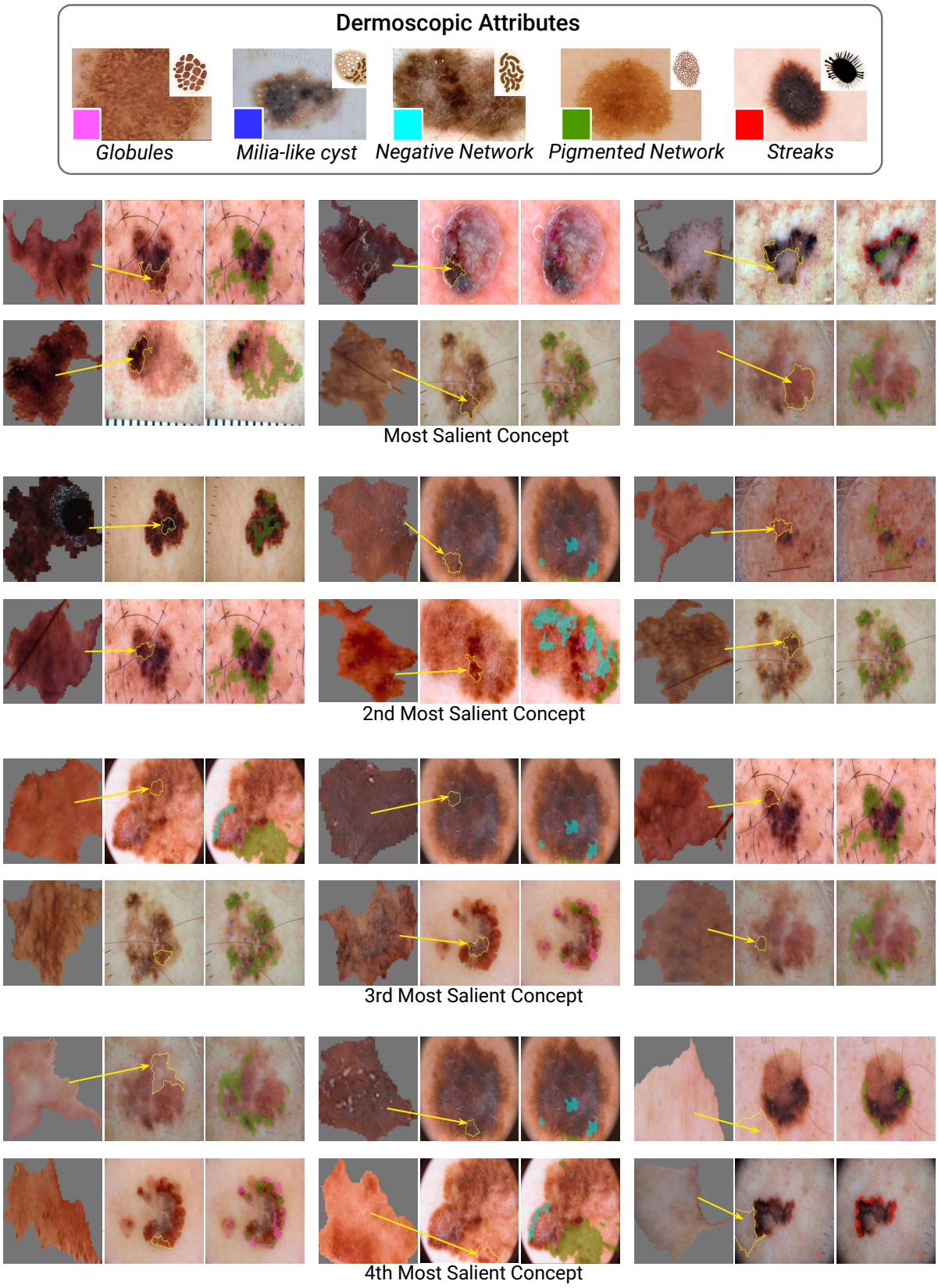


Figure 5.3: Six random examples of the top-4 important concepts from ACE for Inception-v4 in layer *mixed7* for melanoma class. First, the segment is presented, then the position of the segment in the image, and finally, the dermoscopic attributes.

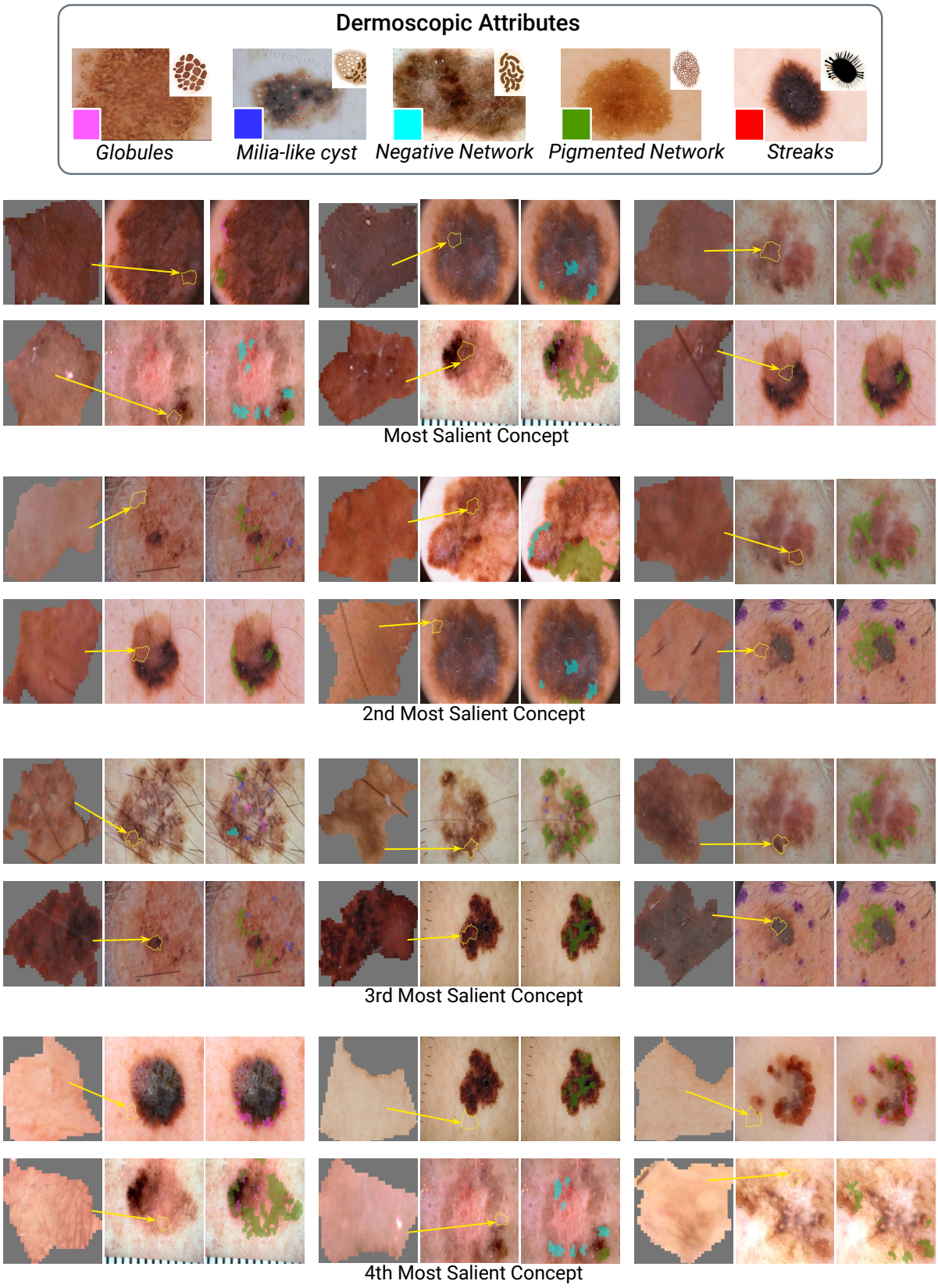


Figure 5.4: Six random examples of the top-4 important concepts from ACE for ResNet-50 in the last convolutional layer for melanoma class. First, the segment is presented, then the position of the segment in the image, and finally, the dermoscopic attributes.

5.3.2 ICE

Figure 5.5 shows ICE results for melanoma class for Inception-v4 and Figure 5.6 for ResNet-50. Similar to ACE, it is hard to interpret the results. We can see the concept samples show lesion parts, but sometimes there is no corresponding dermoscopic attribute, or they are not the same across the samples; therefore, it is difficult to match them with the attributes we have. Furthermore, since this model uses activations and a threshold in the feature map to visualize concepts, the sample concepts are like those obtained with CAM methods, i.e., they show a big part of the lesion without indicating a specific part, being difficult to contrast them with the dermoscopic attributes. As can be seen, concept samples are, in general, very alike between them in color and texture.

Figure 5.7 shows ICE results for the benign class for Inception-v4. The most important concepts are regions outside the lesion; the first concept seems to verify the presence of color patches, the second concept seems to be related to skin hair, and the fourth concept focuses on areas beyond the lesion. For ResNet-50 (Figure 5.8), while the two first more important concepts seem to focus on the lesion, the third seems to check the presence of bubble gels, and the fourth concept is related to surrounding parts of the lesion.

ICE can provide local explanations. Figure 5.9 shows one example of melanoma images predicted correctly (true positive).

In the medical domain, false negative predictions are the most undesired. Figure 5.10 shows the explanation obtained with ICE. As can be seen, the model focuses more on non-lesion parts. The most similar concept is focused on surrounding parts with skin hair, and the third nearest concept seems to be related to the presence of gel bubbles. This indicates that Inception-v4 is basing its decision on spurious correlations to predict a melanoma image as benign, which must not happen in this task.

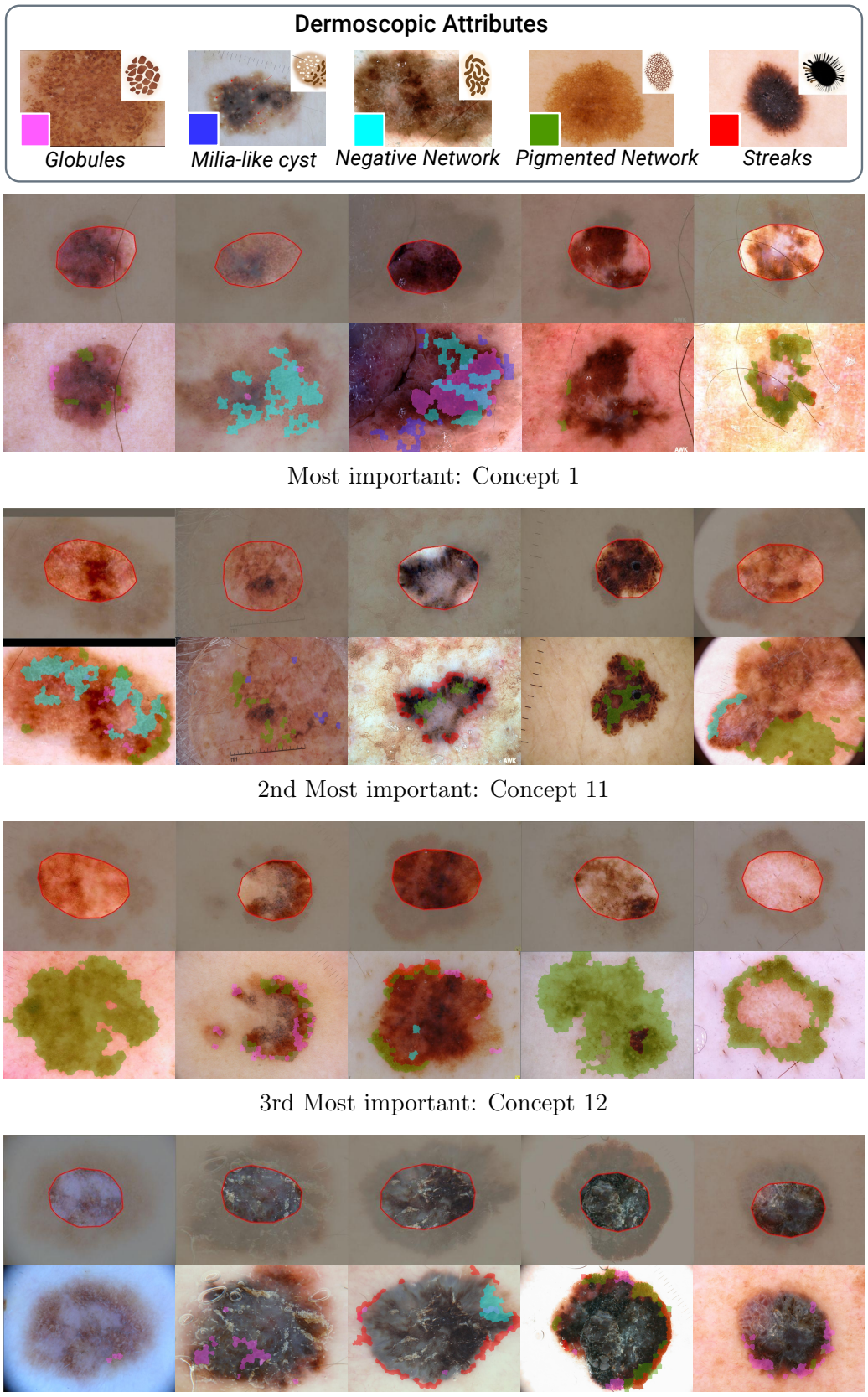


Figure 5.5: Five examples of the top-4 important concepts from ICE in the last layer of Inception-v4 for melanoma class. The first row shows the lesion part related to the concept, and the second row shows dermoscopic attributes.

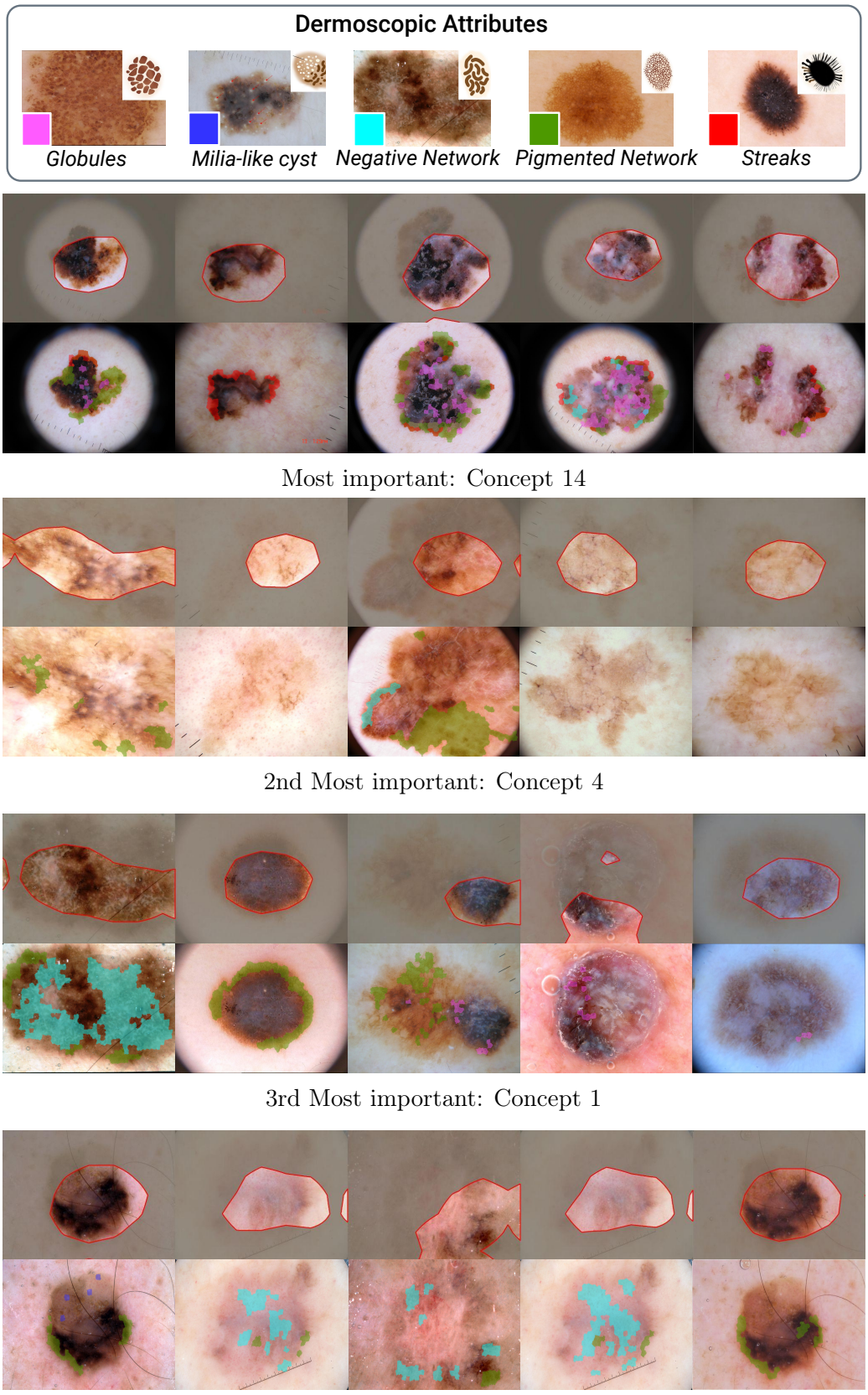


Figure 5.6: Five examples of the top-4 important concepts from ICE in the last layer of ResNet-50 for melanoma class. The first row shows the lesion part related to the concept, and the second row shows dermoscopic attributes.

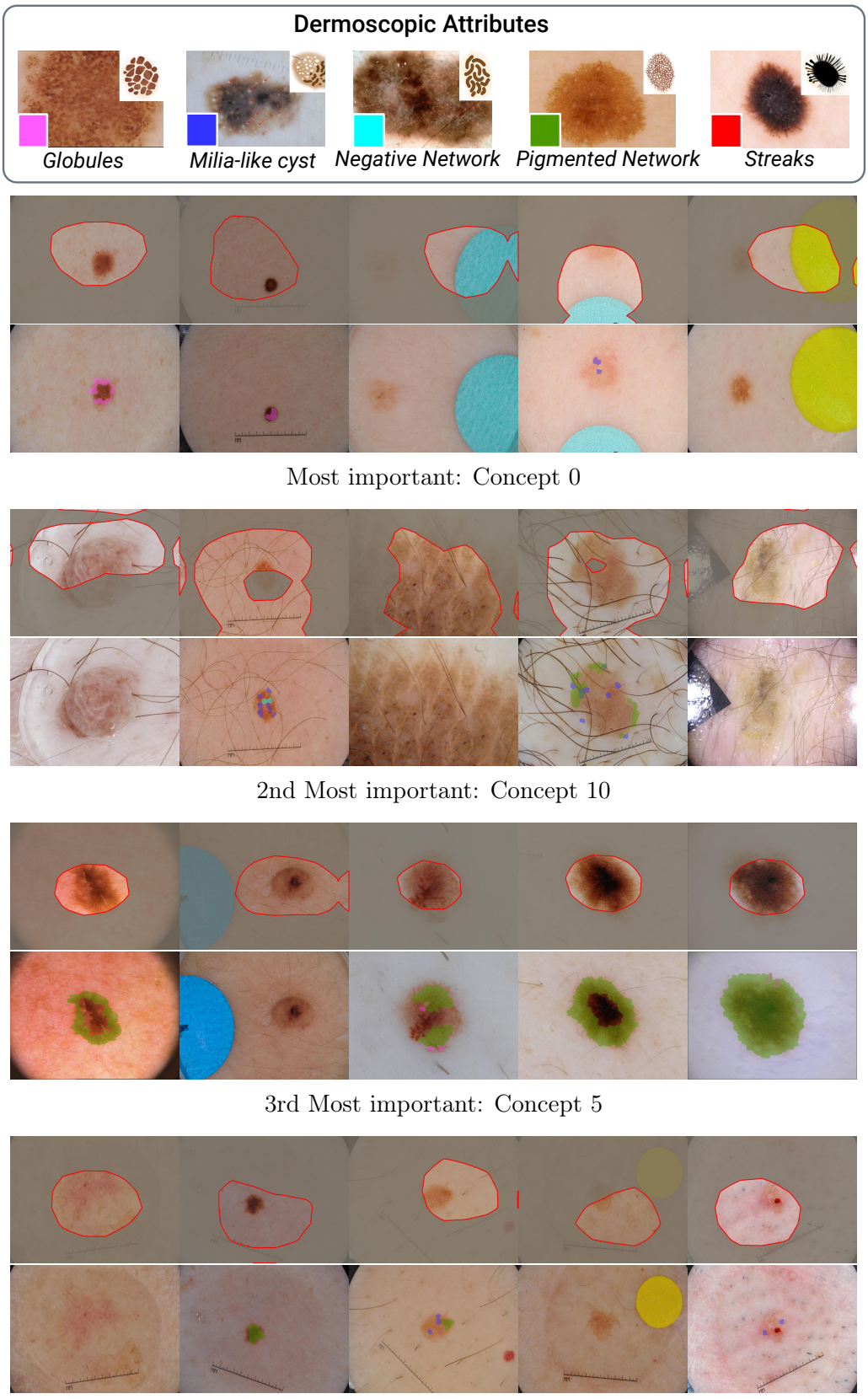


Figure 5.7: Five examples of the top-4 important concepts from ICE in the last layer of Inception-v4 for benign class. The first row shows the lesion part related to the concept, and the second row shows dermoscopic attributes.

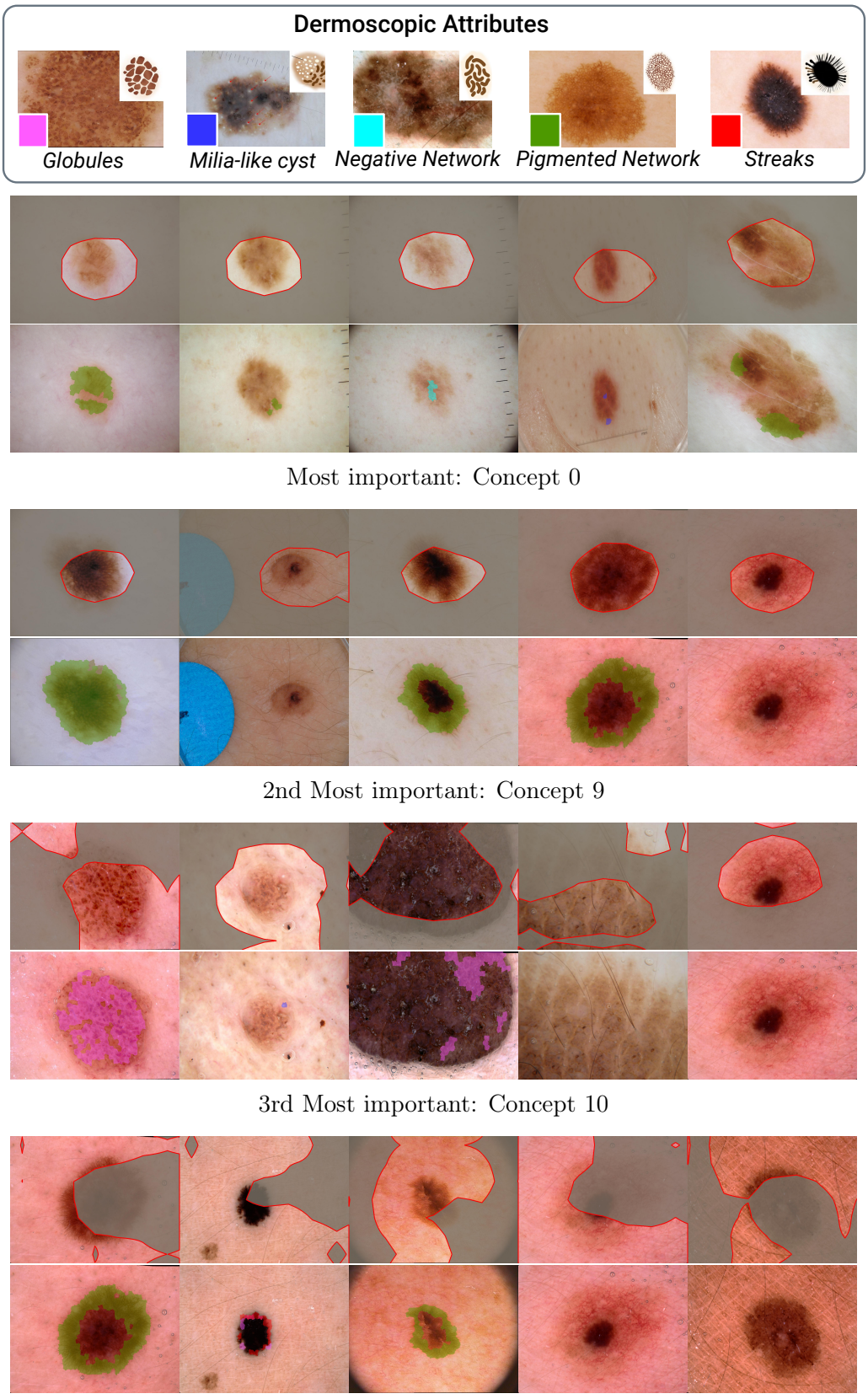


Figure 5.8: Five examples of the top-4 important concepts from ICE in the last layer of ResNet-50 for benign class. The first row shows the lesion part related to the concept, and the second row shows dermoscopic attributes.

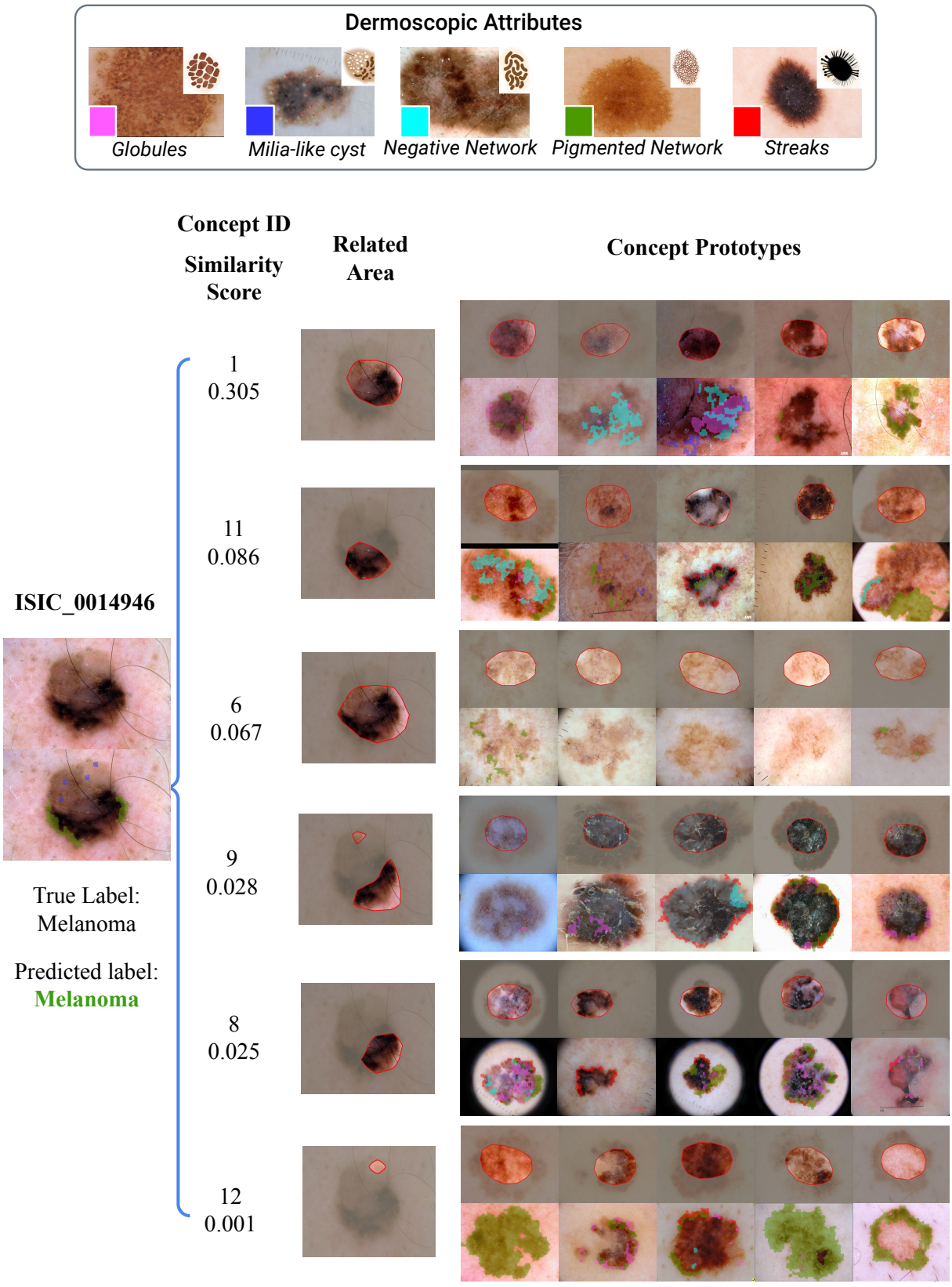


Figure 5.9: Local explanation produced by ICE for ISIC_0014946 – Inception-v4 (true positive).

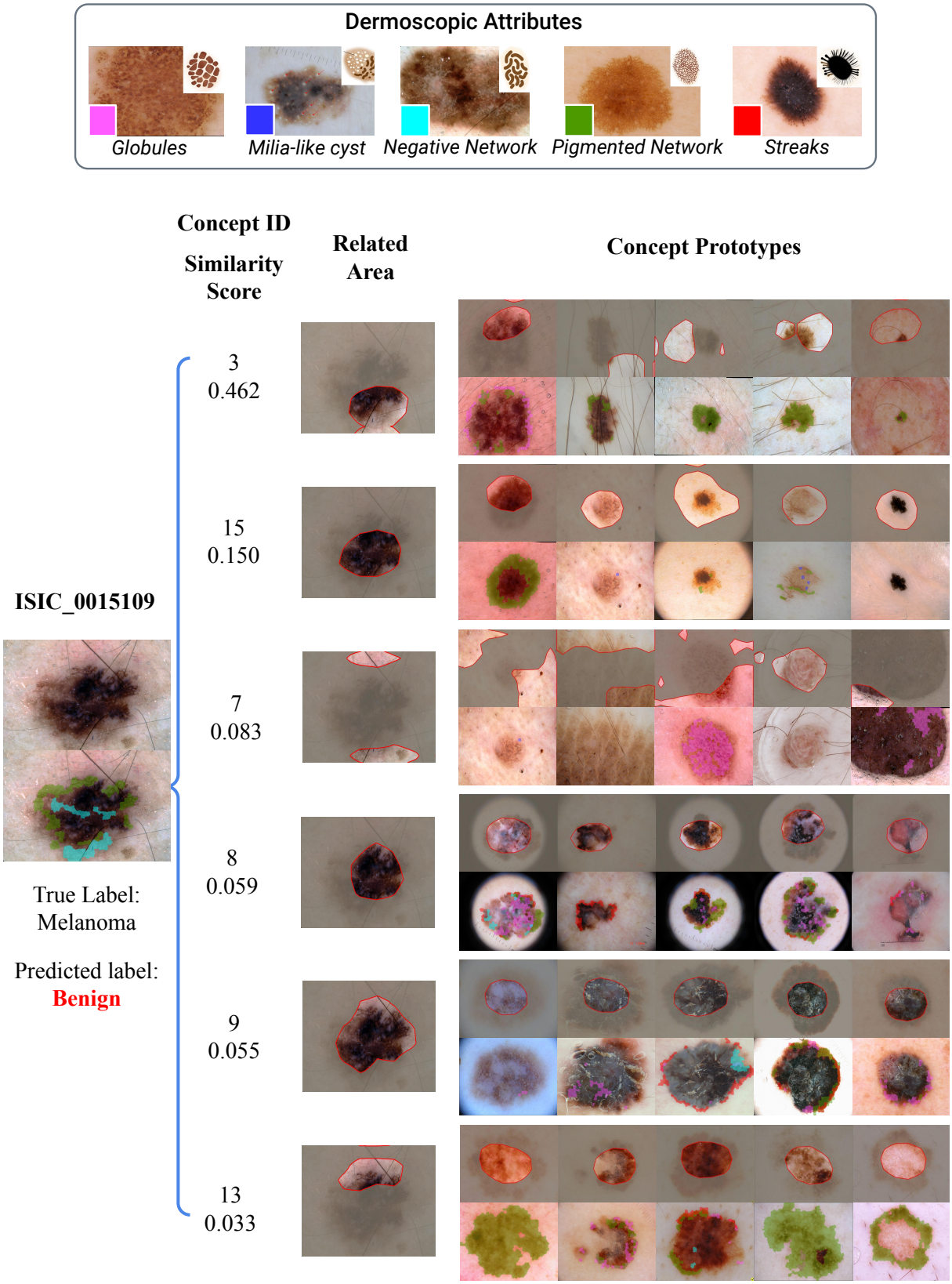


Figure 5.10: Local explanation produced by ICE for ISIC_0015109 – Inception-v4 (false negative).

5.3.3 CME

CME uses model distillation to explain a neural network. First, it maps the layer's activations to concepts, and then, it gets the final output from the concept prediction. We performed different experiments according to the number of concepts.

We used all 25 concepts in the Derm-7pt dataset. Table 5.1 shows the performance results we obtained using CME. We measured fidelity using the predicted labels from the networks Inception-v4 and ResNet-50 as ground truth labels and the task performance using the real ground truth labels. As can be seen, fidelity and task performance results using a Decision Tree (DT) were lower than using Logistic Regression(LR). Besides, given that the Inception-v4 performance to predict melanoma vs. benign images is 0.77 AUC and 0.81 for ResNet-50, we can notice that when using the Logistic Regression, the task performance improved.

Table 5.1: Performance (ROC AUC) of CME extracted models for Inception-v4 and ResNet-50 using all concepts.

	Inception-v4		**ResNet-50**	
	LR	**DT**	**LR**	**DT**
Fidelity of extracted models	0.93	0.88	0.80	0.69
Task performance of extracted models	0.79	0.74	0.85	0.64

Table 5.2 shows the 18 most important concepts from CME global explanation found to predict melanoma. Since it is a logistic regression, we can use the odds to interpret the coefficients in the model. For Inception-v4, the presence of a blue-white veil in the lesion increases the odds of melanoma vs. benign by a factor of 2.96 when all other features remain the same. Alternatively, we can say that: People with a blue-white veil in the lesion have 196% ($2.96 - 1 = 1.96$) more odds of melanoma than those without dermoscopic attributes.

Having the concepts and their importance in Table 5.2, we can hypothesize if the concepts with higher weights are indeed dermoscopic features associated with melanoma according to the International Dermoscopy Society (IDS)[9]. In the first case, for Inception-v4, we can note that most of the 9 highest concepts are associated with melanoma except for two of them: *Vascular structures: hairpin* that can be present in both benign and melanoma lesions and *Vascular structures: comma* that is more associated with nevus (benign) lesions. Also, *Streaks: irregular* was given a negative weight when it should be the contrary. In the second case, for ResNet-50, we can observe a better scenario all 9 highest dermoscopic attributes are associated with melanoma. However, *Regression structures: blue areas* has a negative weight when it should be the contrary.

Figure 5.11 shows the Inception-v4 model distilled into a Decision Tree. We can trace a path in a Decision Tree to understand what rules the model followed to make a particular prediction (see Figure 5.12).

[9] `https://dermoscopedia.org/`

Table 5.2: Concepts weights retrieved from CME: Logistic Regression coefficients. The highest weight indicates more importance for melanoma, and the lowest indicates more importance for benign.

Concept : Dermoscopic Attribute	**Weight**	**Odds**
Regression structures: blue areas	2.802	16.474
Regression structures: white areas	1.686	5.399
Vascular structures: within regression	1.201	3.322
Vascular structures: hairpin	0.731	2.077
Pigment net: atypical	0.458	1.581
Pigmentation: localized irregular	0.348	1.417
Vascular structures: comma	0.184	1.201
Dots globules: irregular	0.137	1.146
Pigmentation: diffuse irregular	0.103	1.108
⋮	⋮	⋮
Vascular structures: dotted	-0.256	0.774
Dots globules: absent	-0.309	0.734
Pigmentation: absent	-0.314	0.731
Streaks: irregular	-0.642	0.526
Streaks: absent	-0.674	0.510
Regression structures: absent	-0.786	0.456
Pigmentation: diffuse regular	-1.142	0.319
Streaks: regular	-1.177	0.308
Dots globules: regular	-2.329	0.097

(a) Results for Inception-v4

Concept : Dermoscopic Attribute	**Weight**	**Odds**
Vascular structures: linear-irregular	1.557	4.747
Dots globules: irregular	0.991	2.694
Streaks: irregular	0.918	2.505
Pigment net: atypical	0.834	2.304
Regression structures: white areas	0.758	2.134
Pigmentation: diffuse irregular	0.648	1.912
Vascular structures: within regression	0.370	1.448
Pigmentation: localized irregular	0.349	1.418
Regression structures: combinations	0.248	1.282
⋮	⋮	⋮
Vascular structures: absent	-0.482	0.617
Pigment net: typical	-0.622	0.537
Vascular structures: hairpin	-0.746	0.474
Regression structures: blue areas	-0.785	0.456
Vascular structures: comma	-0.819	0.441
Regression structures: absent	-0.848	0.428
Pigment net: absent	-0.876	0.417
Streaks: regular	-0.960	0.383
Vascular structures: dotted	-1.709	0.181

(b) Results for ResNet-50

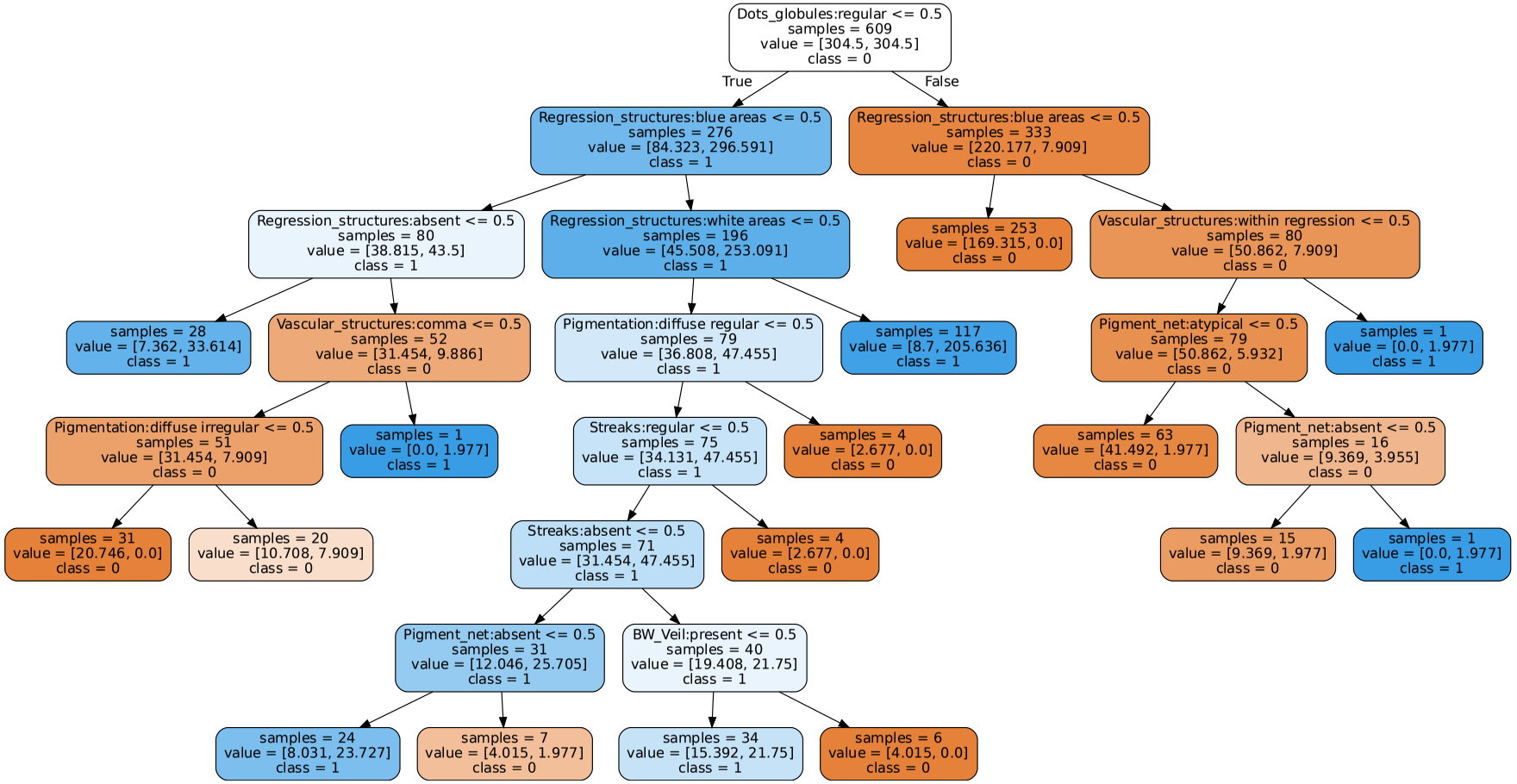


Figure 5.11: Explanation produced by CME for Inception-v4 using Decision Tree. Orange nodes belong to the benign class and blue nodes belong to the melanoma class; the darker the color, the more pure the node.

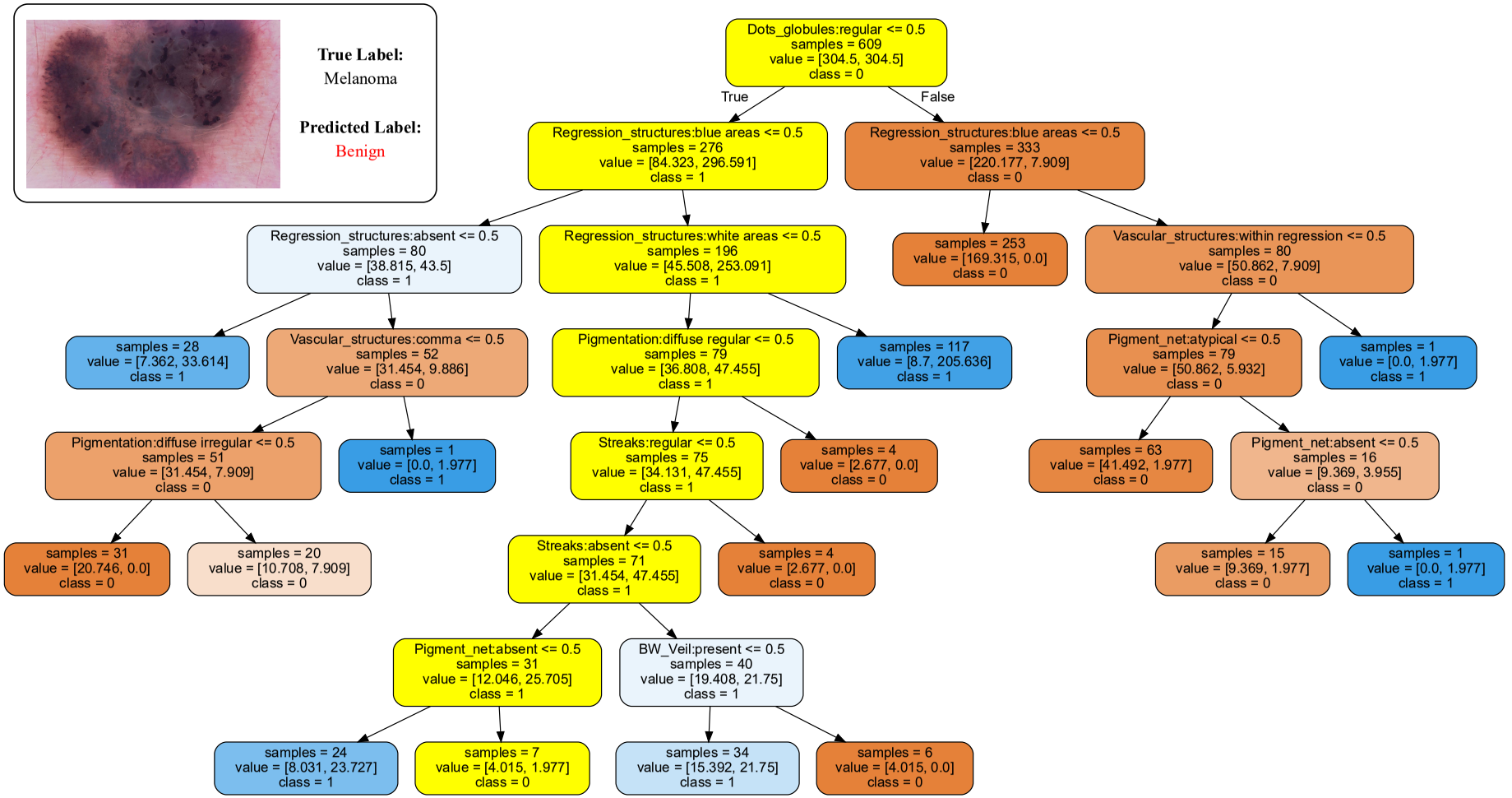


Figure 5.12: Path explaining the prediction for an image produced by CME for Inception-v4 using Decision Tree. Orange nodes belong to the benign class, and blue nodes belong to the melanoma class; the darker the color, the more pure the node.

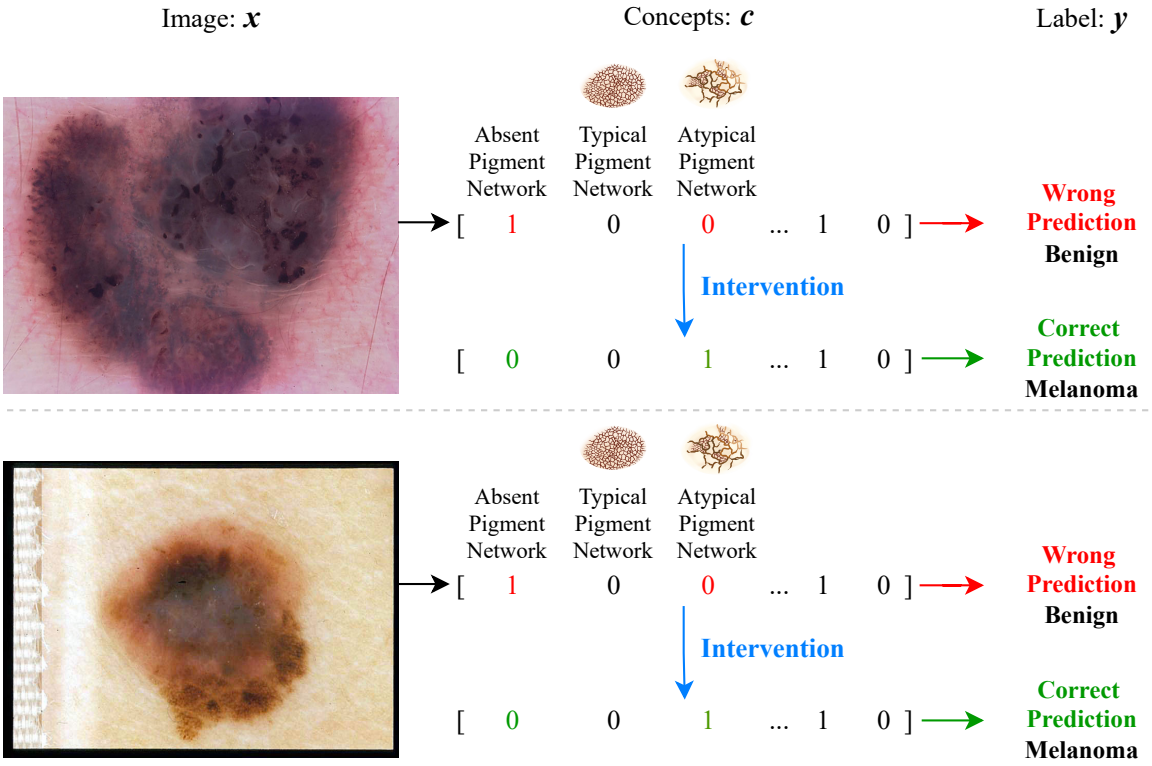


Figure 5.13: Concept intervention on two samples at test time.

We must note that the Concept Predictor got an average of 61.13% ROC AUC for Inception-v4 and 60.5% ROC AUC for ResNet-50. This difference in performance indicates that for the distilled model is more difficult to predict the concepts than to predict melanoma. Therefore, the DNNs trained are not concept-decomposable at all. This can be due to the small number of samples for each concept.

Concept Intervention Recall that we are using the features extracted in the model to predict a concept (dermoscopic attribute), and from there, we get the labels. This way, we can intervene if the model incorrectly predicted the concepts. For example, Figure 5.13 shows two examples of false negative images where the Image-to-Concepts predictor misclassified the attributes related to Pigment Network, and when intervening on them, changing two concepts to its real value, the distilled model gets a correct prediction at test time on both label predictors (Logistic Regression and Decision Tree).

As stated by the authors, an intervention that corrects the distilled model by retraining the concepts-to-label predictor is also possible. For this, we trained a Logistic Regression on the values of the 25 concepts and got the concepts ordered by their coefficient value. Then, we modified the concepts extracted for the train and test set and verified the change in the task performance considering the order. Figure 5.14 shows how the performance changed when modifying the concepts in order of importance. Notice the considerable improvement for the extracted model when only one concept was modified.

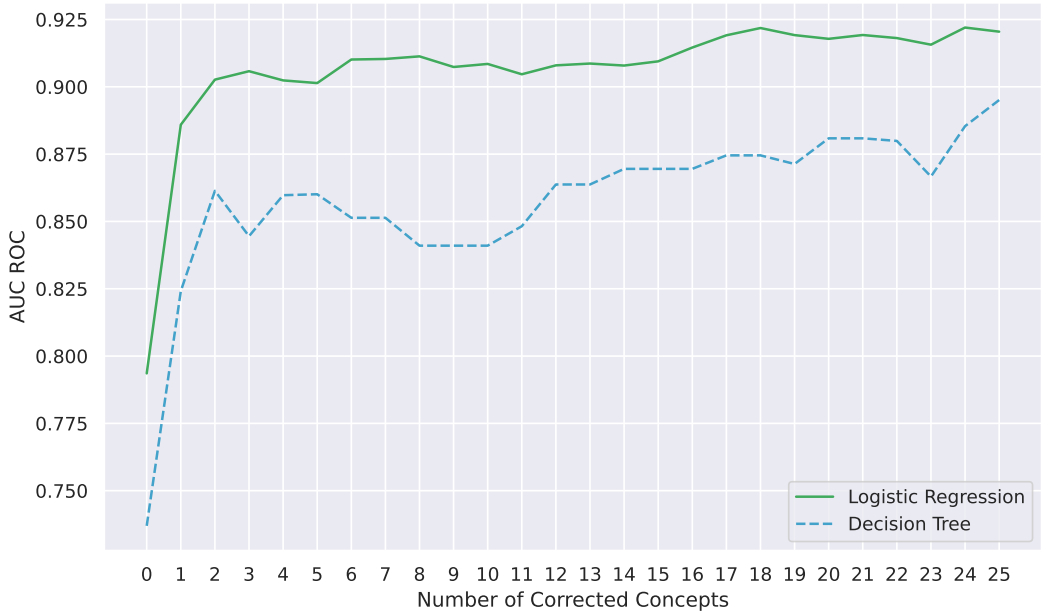
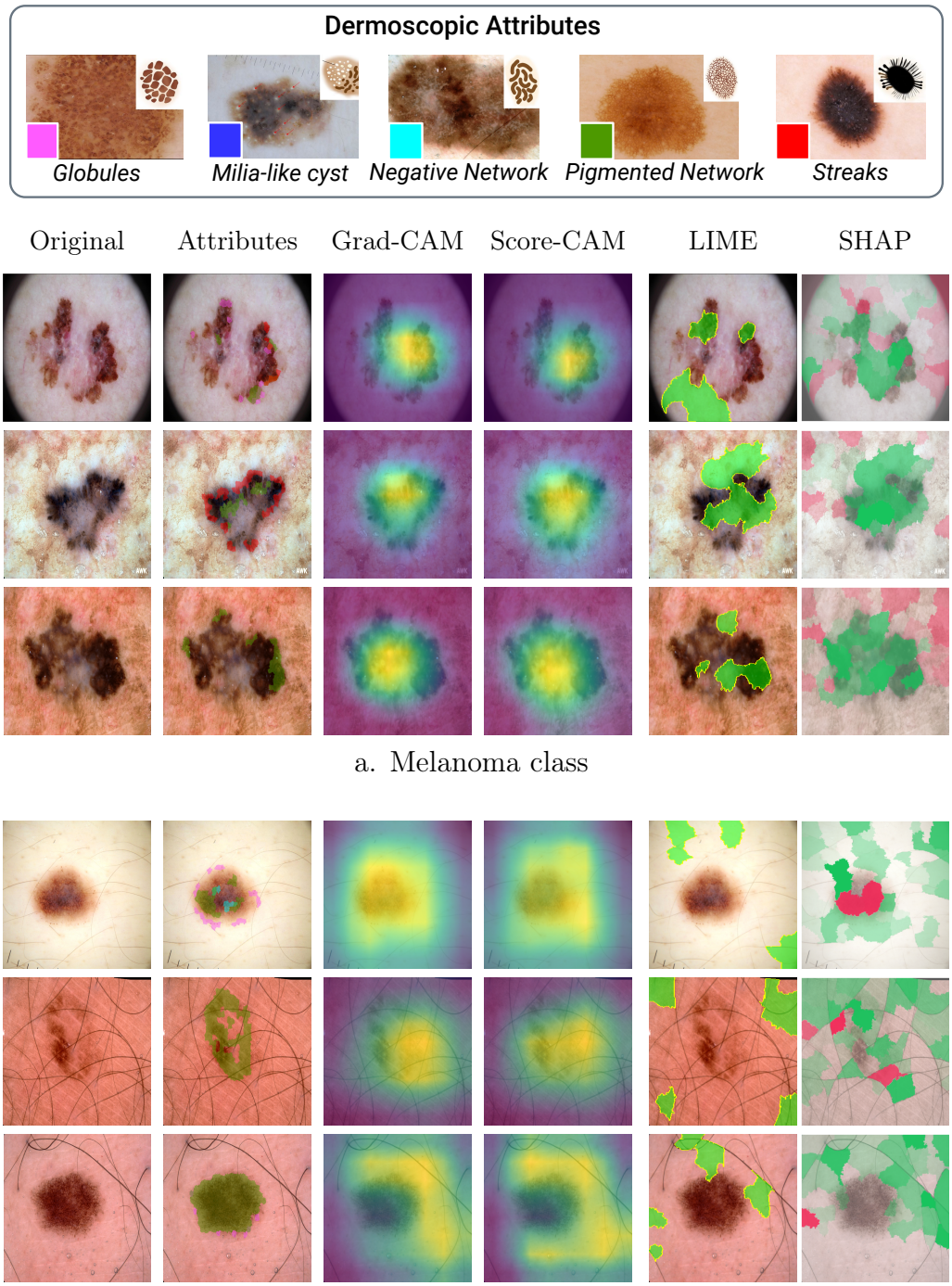


Figure 5.14: Concept intervention on predicted concepts and retraining the label predictor.

5.4 Desiderata Assessment

Regarding the *fidelity*, which measures how much the explanations reflect the real model, we performed a qualitative evaluation following Sun et al. [88] comparing explanations for the predicted images with the highest confidence, and if the methods are explaining accurately the model behavior, the explanations across all methods will be similar. Figure 5.15 and Figure 5.16 show explanations for images with over 99% prediction confidence in both DNNs. For the melanoma class, while Grad-CAM Score-CAM and SHAP are alike, sometimes LIME (see Figure 5.15 a. first row and Figure 5.16 a. third row) does not highlight the same areas the other methods do. For the benign class, where the models got better results in prediction, we observe a similar scenario where LIME and SHAP are getting different explanations than Grad-CAM and Score-CAM (see Figure 5.15 a. first and second row and Figure 5.15 a. second row). Another cause to doubt the correctness of explanations is the fact that the Grad-CAM and Score-CAM saliency maps appear to activate in regions where SHAP has shown to be negatively contributing to the prediction (see Figure 5.15 a. first and second row and Figure 5.15 a. second row).

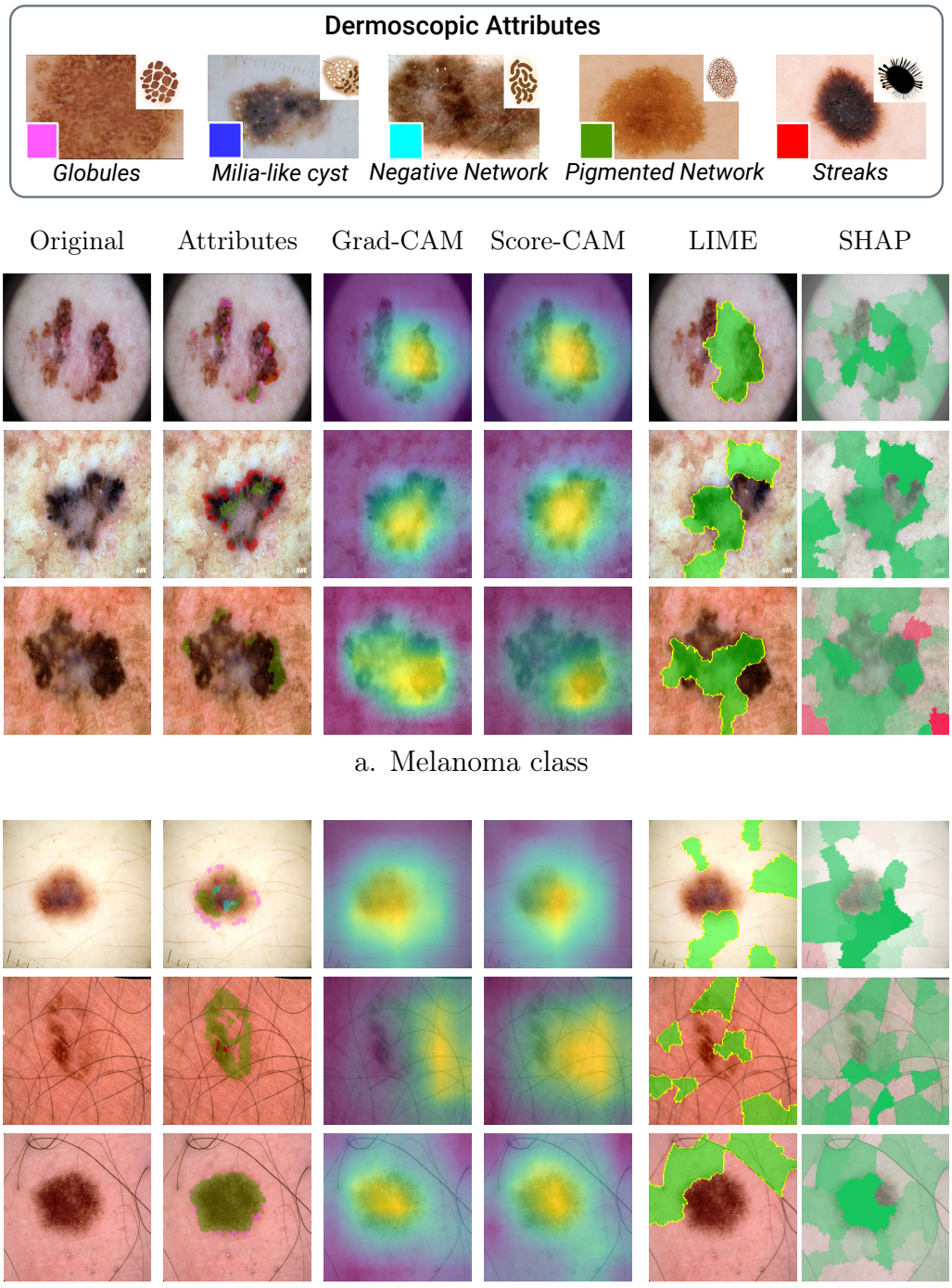
Furthermore, some of the used methods are built to have the highest fidelity possible, such as ICE, based on factorization and distillation methods. For ICE that uses Non-negative matrix factorization, the fidelity error was 11.83% for Inception-v4 and 12.81% for ResNet-50. For CME using Logistic Regression, the fidelity for Inception-v4 was 0.93 ROC AUC and 0.88 ROC AUC for ResNet-50. This result indicates a trade-off between interpretability and the model performance since the methods cannot recover the predictions obtained with the DNN models. Thus, around 10% is getting an incorrect explanation, a considerable percentage for a crucial domain as it is identifying cancer.



a. Melanoma class

b. Benign class

Figure 5.15: Saliency results for predictions with highest confidence with Inception-v4 to test fidelity; ideally, explanation will be similar across all methods. Yellow colors in maps represent the parts of the input images that were more relevant for the prediction. For LIME, green parts represent the image that contributed positively to the prediction. For SHAP, pixels in green are those which contributed positively, and pixels in red are those which contributed negatively.



a. Melanoma class

b. Benign class

Figure 5.16: Saliency results for predictions with highest confidence with ResNet-50 to test fidelity, ideally explanation will be similar across all methods. Yellow colors in maps represent the parts of the input images that were more relevant for the prediction. For LIME, green parts represent the image that contributed positively to the prediction. For SHAP, pixels in green are those which contributed positively, and pixels in red are those which contributed negatively.

Regarding the *meaningfulness* of explanations, concept-based explanations provide more information than pixel-attribution methods and are easier to understand in generic contexts. While CME associates machine parameters with high-level concepts ensuring human comprehension, the comprehensibility of the other concepts-based methods that find concepts in an unsupervised setting (ICE and ACE) requires domain knowledge and can be time-consuming, according to the experience in the domain. Also, there is no guarantee that the concepts found are related to human concepts since the examples of concepts are determined and limited to the superpixel segmentation algorithm (ACE) and the threshold to limit the attention map (ICE).

Concerning the *effectiveness* of explanations, which ensures that the explanation is complete, pixel-attribution methods do not provide enough information to understand why a prediction is being made. They highlight the relevant parts but do not indicate how the model uses them to lead to a particular prediction or what the DNN model is doing with those specific parts of the images. Therefore it is not possible to generate a hypothetical set of rules to be able to simulate the outcome of the model. On the other hand, concept-attribution methods provide more information about how they use the concepts to get the model output.

The two last desired properties are subjective. For this dissertation, we assessed them from a data scientist's point of view. However, when evaluated with other stakeholders (patients, physicians, dermatologists), the appreciation could be different.

5.5 Conclusion

Pixel-attribution methods (Grad-CAM, Score-CAM, LIME, and SHAP) help to ensure that the models work appropriately before deployment by checking the presence of biases and spurious correlations. However, for our scenario (melanoma classification), they do not justify how the DNNs are using the relevant pixels or superpixels.

Since ACE and ICE produce global explanations that identify the model's patterns to predict a particular class, biases, and spurious correlations can be found faster than pixel-attribution methods. Nevertheless, comprehending the concepts, these methods provide as explanation is difficult and requires domain knowledge.

At first sight, CME is a good option for understanding DNNs, given that it uses known high-level concepts to justify the DNN's decision. On the other hand, even a fidelity of over 90% leaves a great chance of doubts about whether to trust the explanation. Furthermore, CME is forcing the DNN's representations to be mapped to the provided human concepts; this could be a limitation to finding new concepts or even biases or spurious correlations. Moreover, more is needed to assess this method with the fidelity obtained; the input-to-concept predictor's performance must also be analyzed since it assures that DNNs are indeed concept-decomposable.

Chapter 6

Conclusion

Explainable Artificial Intelligence (XAI) is vital in deploying responsible AI systems, helping increase trustworthiness and mitigate bias in health care. With the inclusion of a 'right of explanation' in the European General Data Protection Regulation, an AI regulation, many techniques to explain the black-box nature of deep neural networks (DNNs) in a post-hoc manner have been introduced. In this sense, patients and dermatologists need and have the right to know why a DNN is making a particular decision for skin lesion classification.

In this Master's dissertation, we analyzed several articles that used XAI for skin lesion analysis and found that most of them used pixel-attribution methods (saliency methods) superficially, just as verification that the DNN proposed is looking at the right place (skin lesion).

There is still no formalization or consensus in the literature about how to evaluate the XAI methods; each work presents its own test even when they are working on the same task, e.g., image classification using ImageNet (that it is not related to high-stakes tasks, but it is the most common task where XAI methods are demonstrated on). Therefore, we identified three desired properties that explanations should at least accomplish to get transparency and understanding in the DNN decisions for skin lesions classification.

We compared seven state-of-the-art explainability methods, which use different approaches. However, none of them provide a sufficient explanation to understand DNN decisions fully. This problem is critical for deploying automated skin-lesion analysis. When performing in a real-world scenario, we want the network to be a responsible AI for skin-lesion analysis. The explainability methods evaluated in this work still do not — and maybe never will be — provide sufficient information to solve the black-box issue in skin-lesion models [75].

Therefore, the answer for our main research questions proposed in Section 1.4 is:

Q1. Considering that classifying skin lesion images is very different from common domains as ImageNet due to input homogeneity, is it possible to apply current explainability methods to understand skin lesion models?

No, it is not possible to understand the skin lesion models applying current explainability methods. Each method has its advantages and disadvantages. Cur-

rent XAI methods should not be expected to accomplish all goals of explainability. For example, a method designed to find bias and spurious correlations might not work well in assisting decisions makings or increasing trust. Therefore, caution is suggested when using and creating XAI methods, as well as stating end goals and application scenarios that should be considered.

6.1 Future Work

In the AI community, there is a belief about the accuracy-interpretability/explainability trade-off, indicating that the more interpretable or explainable a model is, the less accurate it is. However, this has been demystified by Rudin [75] and Kim [47]. Therefore, the first direction when designing a skin lesion classifier would be to design an explainable model. For example, Abl is a work-in-progress [27] that creates an explainable model using mutual information minimization between human and machine features.

Our work has a main limitation: we only used images from HAM10000 to train two DNN models, Inception-v4 and ResNet-50. Also, we evaluated the explainability methods using data with concept annotations that the models have never seen. Therefore, the models' performance could have been higher than if it were trained with more samples from other datasets like ISIC 2019 and ISIC 2020. Another future direction is to perform more experiments with other architectures; deeper architectures have shown better performance. Recently, a new database [29] of skin diseases and annotations was released, consequently, experiments with this new dataset could be another direction.

Another future direction is a deep study of how different physicians perceive and interpret each explanation. This would bring a broader understanding of the explainability needs in the skin lesion domain.

In this Master's dissertation, we used different methods whose explanations are built upon approximating predictions with an interpretable model. LIME and SHAP, model-agnostic methods, provided local explanations with attributions of superpixels in the image. CME is a model distillation method that maps the feature's activation from the images into textual concepts and uses them to get the same DNN prediction. These three methods optimize the average fidelity of the interpretable models they use. However, we are not considering if the correctness of these explanations changes drastically for a specific group, e.g., groups of different skin tones, gender, and ethnicity. Therefore, a deep study of homogeneity (i.e., fairness in explanations) is needed when evaluating explanations. For example, Balagopalan et al. [8] propose some metrics for this scenario.

6.2 Ethics Statement

The datasets used for this work have an over-representation of lighter skin tones. This bias has not only affected the DNN models used in this research but could be perpetuated by the explanations. Also, having post-hoc explanations can lead to a false sense of understanding and confidence in AI models. XAI should not be seen as an escape from accountability but as a step towards a responsible deployment of AI [53].

Bibliography

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012. 49

[2] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018. 15

[3] I. A. Alfi, M. Mahfuzur Rahman, M. Shorfuzzaman, and A. Nazir. A non-invasive interpretable diagnosis of melanoma skin cancer using deep learning and ensemble stacking of machine learning models. *Diagnostics*, 12, 2022. 39

[4] S. Allegretti, F. Bolelli, F. Pollastri, S. Longhitano, G. Pellacani, and C. Grana. Supporting skin lesion diagnosis with content-based image retrieval. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8053–8060, 2021. 37

[5] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020. 14, 18, 19, 21

[6] N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J. Patel, M. Gidwani, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6), 2021. 15

[7] S. Bach, A. Binder, G. Montavon, F.rederick Klauschen, K. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), 2015. 22

[8] A. Balagopalan, H. Zhang, K. Hamidieh, T. Hartvigsen, F. Rudzicz, and M. Ghassemi. The road to explainability is paved with bias: Measuring the fairness of explanations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, page 1194–1206, 2022. 73

[9] C. Barata, M. E. Celebi, and J. S. Marques. Explainable skin lesion diagnosis using taxonomies. *Pattern Recognition*, 110, 2021. 37

[10] C. Barata and C. Santiago. Improving the explainability of skin cancer diagnosis using CBIR. In *Medical Image Computing and Computer Assisted Intervention*, pages 550–559, 2021. 37

[11] W. Barhoumi and A. Khelifa. Skin lesion image retrieval using transfer learning-based approach for query-driven distance recommendation. *Computers in Biology and Medicine*, 137:104825, 2021. 37

[12] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 28

[13] O. Biran and C. Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI*, volume 8, pages 8–13, 2017. 18

[14] A. Bissoto, C. Barata, E. Valle, and S. Avila. Artifact-based domain generalization of skin lesion models. In *Computer Vision–ECCV 2022 Workshops*, pages 133–149, 2022. 14

[15] A. Bissoto, M. Fornaciali, E. Valle, and S. Avila. (De)constructing bias on skin lesion datasets. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 50

[16] A. Bissoto, F. Perez, E. Valle, and S. Avila. Skin lesion synthesis with generative adversarial networks. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 294–302, 2018. 14

[17] A. Bissoto, E. Valle, and S. Avila. Debiasing skin lesion datasets and models? Not so fast. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 50

[18] A. Bissoto, E. Valle, and S. Avila. Gan-based data augmentation and anonymization for skin-lesion analysis: A critical review. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1847–1856, 2021. 14

[19] A. Boggust, B. Hoover, A. Satyanarayan, and H. Strobelt. Shared interest: Measuring human-ai alignment to identify recurring patterns in model behavior. In *Conference on Human Factors in Computing Systems*, 2022. 39

[20] T. J. Brinker, A. Hekler, A. H. Enk, J. Klode, A. Hauschild, C. Berking, B. Schilling, S. Haferkamp, D. Schadendorf, S. Fröhling, J. S. Utikal, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer*, 111:148–154, 2019. 16

[21] O. M. Camburu. *Explaining deep neural networks*. PhD thesis, University of Oxford. 19

[22] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 2019. 42

[23] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Conference on Applications of Computer Vision*, pages 839–847, 2018. 22

[24] L. Chaves, A. Bissoto, E. Valle, and S. Avila. An evaluation of self-supervised pre-training for skin-lesion analysis. In *Computer Vision–ECCV 2022 Workshops*, pages 150–166, 2022. 14

[25] Z. Chen, Y. Bei, and C. Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020. 46

[26] T. Chowdhury, A. R. S. Bajwa, T. Chakraborti, J. Rittscher, and U. Pal. Exploring the correlation between deep learned and clinical features in melanoma detection. In *Medical Image Understanding and Analysis: Annual Conference*, pages 3–17, 2021. 38

[27] E. Cobos, T. Kuestner, B. Schölkopf, and S. Gatidis. Explainable medical image analysis by leveraging human-interpretable features through mutual information minimization. In *Medical Imaging meets NeurIPS 2021 Workshop at Conference on Neural Information Processing Systems*, 2021. 73

[28] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2019. 43

[29] R. Daneshjou, M. Yuksekgonul, Z. Ran Cai, R. A Novoa, and J. Zou. Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 73

[30] A. Das and P. Rad. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *arXiv*, pages 1–24, 2020. 18

[31] F. Eitel and K. Ritter. Testing the robustness of attribution methods for convolutional neural networks in mri-based alzheimer's disease classification. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pages 3–11, 2019. 15

[32] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009. 22

[33] R. Fong, M. Patrick, and A. Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *IEEE International Conference on Computer Vision*, pages 2950–2958, 2019. 22

[34] R. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *IEEE International Conference on Computer Vision*, pages 3429–3437, 2017. 22

[35] A. Ghandeharioun, B. Kim, C.-L. Li, B. Jou, B. Eoff, and R. W. Picard. Dissect: Disentangled simultaneous explanations via concept traversals. *arXiv preprint arXiv:2105.15164*, 2021. 37

[36] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, pages 9277–9286, 2019. 29, 30, 46, 47, 48, 49

[37] B. Goodman and S. Flaxman. Eu regulations on algorithmic decision-making and a "right to explanation". In *ICML Workshop on Human Interpretability in Machine Learning*, 2016. 19

[38] S. Haggenmüller, R. C. Maron, A. Hekler, J. S. Utikal, C. Barata, R. L. Barnhill, H. Beltraminelli, C. Berking, et al. Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. *European Journal of Cancer*, 156:202–216, 2021. 16

[39] M. Harradon, J. Druce, and B. Ruttenberg. Causal learning and explanation of deep neural networks via autoencoded activations. *arXiv preprint arXiv:1802.00541*, 2018. 21

[40] K. Hauser, A. Kurz, S. Haggenmüller, R. C. Maron, C. von Kalle, J. S. Utikal, F. Meier, S. Hobelsberger, F. F. Gellrich, M. Sergon, et al. Explainable artificial intelligence in skin cancer recognition: A systematic review. *European Journal of Cancer*, 167:54–69, 2022. 34

[41] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer, 2016. 19

[42] Kai Huang, Xiaoyu He, Zhentao Jin, Lisha Wu, Xinyu Zhao, Zhe Wu, Xian Wu, Yang Xie, Miaojian Wan, Fangfang Li, et al. Assistant diagnosis of basal cell carcinoma and seborrheic keratosis in chinese population using convolutional neural network. *Journal of healthcare engineering*, 2020, 2020. 38

[43] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2019. 35, 44

[44] Dmitry Kazhdan, Botty Dimanov, Mateja Jamnik, Pietro Liò, and Adrian Weller. Now you see me (cme): Concept-based model extraction. volume 2699, 2020. 32, 33, 46, 47, 48

[45] Dmitry Kazhdan, Botty Dimanov, Helena Andres Terre, Mateja Jamnik, Pietro Liò, and Adrian Weller. Is disentanglement all you need? comparing concept-based & disentanglement approaches. *ICLR 2021 Workshop on Robust and Reliable ML in the Real World*, 2021. 33, 42

[46] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, James. Wexler, F. Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, pages 2668–2677, 2018. 21, 28, 29, 46

[47] Been Kim. Beyond interpretability: developing a language to shape our relationships with ai, 2022. 73

[48] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 2280–2288. Curran Associates, Inc., 2016. 18

[49] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020. 46

[50] Dominika Kwiatkowska, Piotr Kluska, and Adam Reich. Convolutional neural networks for the detection of malignant melanoma in dermoscopy images. *Advances in Dermatology and Allergology/Postepy Dermatol Alergol.*, 38(3):412, 2021. 38

[51] Samira Lafraxo, Mohamed El Ansari, and Said Charfi. Melanet: an effective deep learning framework for melanoma detection using dermoscopic images. *Multimedia Tools and Applications*, 81, 2022. 39

[52] Weipeng Li, Jiaxin Zhuang, Ruixuan Wang, Jianguo Zhang, and Wei-Shi Zheng. Fusing metadata and dermoscopy images for skin disease diagnosis. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1996–2000, 2020. 37

[53] Gabriel Lima, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. The conflict between explainable and accountable decision-making algorithms. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2103–2113, New York, NY, USA, 2022. Association for Computing Machinery. 73

[54] Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In *International Joint Conference on Neural Networks*, pages 1–10, 2020. 35, 40

[55] Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. Exaid: A multimodal explanation framework for computer-aided diagnosis of skin lesions. *Computer Methods and Programs in Biomedicine*, 215, 2022. 35, 40

[56] Scott Lundberg. Shap github repository, 2017. 27

[57] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774. Curran Associates, Inc., 2017. 20, 22, 26, 27, 47, 48

[58] Andrei Margeloiu, Nikola Simidjievski, Mateja Jamnik, and Adrian Weller. Improving interpretability in medical imaging diagnosis using adversarial training. *NeurIPS 2020 workshop Medical Imaging meets NeurIPS (MED-NEURIPS)*, 2020. 37

[59] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. Ph^2-a dermoscopic image database for research and benchmarking. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5437–5440. IEEE, 2013. 35

[60] A. Menegola, M. Fornaciali, R. Pires, F. V. Bittencourt, S. Avila, and E. Valle. Knowledge transfer for melanoma screening with deep learning. In *IEEE International Symposium on Biomedical Imaging*, pages 297–300, 2017. 14

[61] Carlo Metta, Riccardo Guidotti, Yuan Yin, Patrick Gallinari, and Salvatore Rinzivillo. Exemplars and counterexemplars explanations for image classifiers, targeting skin lesion labeling. volume 2021-September, 2021. 37

[62] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1 – 38, 2019. 18

[63] David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *BMJ*, 339, 2009. 34

[64] C. Molnar. *Interpretable Machine Learning*. 2019. `https://christophm.github.io/interpretable-ml-book/`. 18, 26, 28, 42

[65] Fabrizio Nunnari, Md Abdul Kadir, and Daniel Sonntag. On the overlap between grad-cam saliency maps and explainable visual features in skin cancer images. volume 12844 LNCS, 2021. 38

[66] F. Perez, S. Avila, and E. Valle. Solo or ensemble? choosing a cnn architecture for melanoma classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 14

[67] F. Perez, C. Vasconcelos, S. Avila, and E. Valle. Data augmentation for skin lesion analysis. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 303–311, 2018. 14

[68] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and S. S. Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv.*, 51(5), September 2018. 14

[69] Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, and Olga Russakovsky. Overlooked factors in concept-based explanations: Dataset choice, concept salience, and human capability, 2022. 35

[70] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Local interpretable model-agnostic explanations (lime): An introduction, 2016. 26

[71] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1135–1144, 2016. 20, 22, 25, 47, 48

[72] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. *AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. 20

[73] L. Rieger, C. Singh, W. J. Murdoch, and B. Yu. Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. *International Conference on Machine Learning*, PartF16814:8086–8096, 2020. 37

[74] M. Robnik-Šikonja and M. Bohanec. *Perturbation-Based Explanations of Prediction Models*, pages 159–175. 2018. 42

[75] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. 72, 73

[76] Berkman Sahiner, Aria Pezeshk, Lubomir M. Hadjiiski, Xiaosong Wang, Karen Drukker, Kenny H. Cha, Ronald M. Summers, and Maryellen L. Giger. Deep learning in medical imaging and radiation therapy. *Medical Physics*, 46(1):e1–e36, 2019. 14

[77] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus Robert Müller. Toward interpretable machine learning: Transparent deep neural networks and beyond. *arXiv*, pages 1–24, 2020. 19, 21

[78] Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019. 42

[79] Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *arXiv e-prints*, pages arXiv–2105, 2021. 42

[80] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, pages 618–626, 2017. 14, 20, 22, 23, 47, 48

[81] Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953. 26

[82] Sumeet Shinde, Priyanka Tupe-Waghmare, Tanay Chougule, Jitender Saini, and Madhura Ingalhalikar. Predictive and discriminative localization of pathology using high resolution class activation maps with CNNs. *PeerJ Computer Science*, 7:1–14, 2021. 35, 39

[83] M. Shorfuzzaman. An explainable stacked ensemble of deep learning models for improved melanoma skin cancer detection. *Multimedia Systems*, 2021. 39

[84] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. JMLR. org, 2017. 20, 22, 28

[85] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2020. 26

[86] American Cancer Society. Melanoma survival rates | melanoma survival statistics. `https://www.cancer.org/cancer/melanoma-skin-cancer/detection-diagnosis-staging/survival-rates-for-melanoma-skin-cancer-by-stage.html#references`. (Accessed on 11/15/2022). 15

[87] F. Stieler, F. Rabe, and B. Bauer. Towards domain-specific explainable ai: Model interpretation of a skin image classifier using a human approach. 2021. 35, 40

[88] Jia Sun, Tapabrata Chakraborti, and J. Alison Noble. A comparative study of explainer modules applied to automated skin lesion classification. *CEUR Workshop Proceedings*, 2796, 2020. 35, 38, 68

[89] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Gradients of counterfactuals. *arXiv preprint arXiv:1611.02639*, 2016. 20, 22

[90] Matus Telgarsky. Benefits of depth in neural networks. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1517–1539, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. 19

[91] The Global Cancer Observatory, International Agency for Research on Cancer, World Health Organization. All cancers fact sheet. `https://gco.iarc.fr/today/data/factsheets/cancers/39-All-cancers-fact-sheet.pdf`, 2020. 15

[92] The Global Cancer Observatory, International Agency for Research on Cancer, World Health Organization. Melanoma of skin fact sheet. `https://gco.iarc.fr/today/data/factsheets/cancers/16-Melanoma-of-skin-fact-sheet.pdf`, 2020. 15

[93] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813, 2020. 15, 16

[94] P. Tschandl. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, 2018. 43

[95] P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, 2020. 37

[96] P. Tschandl, C. Rosendahl, and H. Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 44

[97] M. Tziomaka and I. Maglogiannis. Ensembles of deep convolutional neural networks for detecting melanoma in dermoscopy images. volume 12876 LNAI, 2021. 39

[98] E. Valle, M. Fornaciali, A. Menegola, J. Tavares, F. V. Bittencourt, L. T. Li, and S. Avila. Data, depth, and design: Learning reliable models for skin lesion analysis. *Neurocomputing*, 2020. 14, 45

[99] Bas HM van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, page 102470, 2022. 15

[100] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds-200-2011 (cub-200-2011). Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 33

[101] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 24–25, 2020. 22, 23, 24, 47, 48

[102] Sutong Wang, Yunqiang Yin, Dujuan Wang, Yanzhang Wang, and Yaochu Jin. Interpretability-Based Multimodal Convolutional Neural Networks for Skin Lesion Diagnosis. *IEEE Transactions on Cybernetics*, pages 1–15, 2021. 39

[103] Lisheng Wei, Kun Ding, and Huosheng Hu. Automatic Skin Cancer Detection in Dermoscopy Images Based on Ensemble Lightweight Deep Learning Network. *IEEE Access*, 8:99633–99647, 2020. 38

[104] Z. Wei, Q. Li, and H. Song. Dual attention based network for skin lesion classification with auxiliary learning. *Biomedical Signal Processing and Control*, 74, 2022. 39

[105] World Health Organization. *Guide to cancer early diagnosis*. 2017. 14, 15

[106] Ning Xie, Gabrielle Ras, Marcel van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated. *arXiv preprint arXiv:2004.14545*, 2020. 19, 21, 38

[107] Mengting Xu, Tao Zhang, Zhongnian Li, Mingxia Liu, and Daoqiang Zhang. Towards evaluating the robustness of deep diagnostic models by adversarial attack. *Medical Image Analysis*, 69, 2021. 39

[108] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In *Advances in Neural Information Processing Systems*, 2020. 46

[109] M. Yuksekgonul, M. Wang, and J. Zou. Post-hoc concept bottleneck models. *ICLR 2022 Workshop on PAIR2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, 2022. 35, 40

[110] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833, 2014. 22

[111] Matthew D Zeiler, Graham W Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *International Conference on Computer Vision*, pages 2018–2025, 2011. 22

[112] Quanshi Zhang, Ruiming Cao, Feng Shi, Ying Nian Wu, and Song-Chun Zhu. Interpreting cnn knowledge via an explanatory graph. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 21

[113] Quanshi Zhang, Yu Yang, Haotian Ma, and Ying Nian Wu. Interpreting cnns via decision trees. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6261–6270, 2019. 21

[114] R. Zhang, P. Madumal, T. Miller, K. A. Ehinger, and B. I. P. Rubinstein. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 11682–11690, 2021. 46, 47, 49

[115] Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A Survey on Neural Network Interpretability. X(X):1–16, 2020. 19, 30, 31, 48

[116] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 22

[117] Hasib Zunair and A. Ben Hamza. Melanoma detection using adversarial training and deep transfer learning. *Physics in Medicine and Biology*, 65(13), 2020. 38