Universidade de Brasília – UnB

Faculdade UnB Gama – FGA

Software Engineering

Explainable Artificial Intelligence: Analysing Skin Lesions Classifiers

Author: Matheus Henrique Sousa Costa

Supervisor: Prof. Dr. Nilton Correia da Silva

Brasilia, DF, Brazil

2021

Matheus Henrique Sousa Costa

Explainable Artificial Intelligence: Analysing Skin Lesions Classifiers

Work submitted to the undergraduate course in Software Engineering of the University of Brasília, as a partial requirement to obtain a Bachelor of Software Engineering Degree.

Universidade de Brasília – UnB

Faculdade UnB Gama – FGA

Supervisor: Prof. Dr. Nilton Correia da Silva

Brasilia, DF, Brazil
2021

Matheus Henrique Sousa Costa

Explainable Artificial Intelligence: Analysing Skin Lesions Classifiers/ Matheus Henrique Sousa Costa. – Brasilia, DF, Brazil, 2021-

71 p. : il. (some color.) ; 30 cm.

Supervisor: Prof. Dr. Nilton Correia da Silva

Final Bachelor's Project – Universidade de Brasília – UnB
Faculdade UnB Gama – FGA , 2021.

1. XAI. 2. Deep Learning. I. Prof. Dr. Nilton Correia da Silva. II. Universidade de Brasília. III. Faculdade UnB Gama. IV. Explainable Artificial Intelligence: Analysing Skin Lesions Classifiers

CDU 02:141:005.6

Matheus Henrique Sousa Costa

Explainable Artificial Intelligence: Analysing Skin Lesions Classifiers

Work submitted to the undergraduate course in Software Engineering of the University of Brasília, as a partial requirement to obtain a Bachelor of Software Engineering Degree.

Final Project Approved. Brasilia, DF, Brazil, May 27th, 2021:

Prof. Dr. Nilton Correia da Silva
Supervisor

Prof. Dr. Fabricio Ataides Braz
Invitee 1

Jerônimo da Silva Avelar Filho
Invitee 2

Brasilia, DF, Brazil
2021

Acknowledgements

I am very grateful to all my friends and family who supported me in this academic trajectory. Especially my future wife Juliana Marques, who supported me from the beginning of the course in all possible ways, for her support in good times and bad times; I am grateful to my parents who have always supported me and allowed me to follow my dreams of being a software engineer, for teaching me to become what I am today and for always being so proud of me; to my aunt Rosana, for being an excellent person and teacher, who with love and care provided support with the review of this work; to my great friend Elmar Roberto, for also giving me all support over the years, all friendship that we have transcended beyond the university; to my friends Gabriel Viana, João Almeida, Ramon Sales, Matheus Oliveira and Matheus Barbosa, friendships that the university provide me, for always believing in me and teaching me a lot; to professors Carla Rocha, Elaine Venson, Tiago Alves, Edson Alves and Renato Sampaio who I am extremely grateful for everything they taught me at this time; and of course, to Professor Nilton Silva, who granted me the honor of having him as an advisor, for the trust, support, knowledge and for understanding my difficulties during the period of this final work.

Thank you all, I will always keep you in my heart!

“I’m a shooting star leaping through the sky
Like a tiger defying the laws of gravity
I’m a racing car passing by like Lady Godiva
I’m gonna go, go, go
There’s no stopping me“
(Freddie Mercury, Queen)

Abstract

Considering the severity of the advance of diseases and dermatological lesions such as skin cancer, as well as its various physical manifestations and other implications, the present research had as its central objective the development of an AI using the methodology Explainable Artificial Intelligence (XAI) posthoc on a Convolutional Neural Network (CNN). The use of XAI was due to bringing a greater capacity for interpretability of data in Deep Learning (DL), rigor and ethical commitment in the construction of the model, which was intended to collaborate in the diagnosis of skin diseases when using as Area Under the Curve (AUC) metrics, Infidelity and Sensitivity for comparing methods XAI. The XAI Integrated Gradients, DeepLIFT, DeepSHAP, GradientSHAP, Occusion and GradCAM methods were compared. The results showed that the Integrated Gradients and DeepLIFT had lower Infidelity and Sensitivity in ResNet-152 model using the dataset HAM10000 with skin lesions images. The data obtained were compared with three authors of similar studies in the literature.

Keywords: Artificial Intelligence (AI); Explainable Artificial Intelligence (XAI); Machine Learning (ML); Skin lesions; DeepLearning (DL); Convolutional Neural Networks (CNN).

Resumo

Considerando a gravidade do avanço de doenças e lesões dermatológicas como o câncer de pele, bem como suas diversas manifestações físicas e demais implicações, a presente pesquisa teve como objetivo central o desenvolvimento de uma Inteligência Artificial (IA) utilizando a metodologia Inteligência Artificial Explanável (XAI) posthoc em uma Rede Neural Convolucional (CNN). A utilização de XAI se deu em função de trazer uma maior capacidade de interpretabilidade dos dados no Aprendizado Profundo, rigor e compromisso ético na construção do modelo, que teve como intenção colaborar no diagnóstico de doenças de pele ao utilizar como métricas Área Sob a Curva (AUC), Infidelidade e Sensitividade para comparação dos métodos XAI. Foram comparados os métodos XAI Integrated Gradients, DeepLIFT, DeepSHAP, GradientSHAP, Occusion e GradCAM. Os resultados demonstraram que o Integrated Gradients e o DeepLIFT tiveram menores Infidelidades e Sensitividades no modelo ResNet-152 utilizando o dataset HAM10000 com imagens de lesões de peles. Os dados obtidos foram comparados com três autores de trabalhos similares encontrados na literatura.

Palavras-chave: Inteligência Artificial (IA); Inteligência Artificial Explanável (XAI); Aprendizado de Máquina; Lesões de pele; Aprendizado Profundo; Redes Neurais Convolucionais (CNN).

List of figures

List of tables

Acronyms

AI Artificial Intelligence. 9, 25, 39, 67

AKIEC Actinic keratoses and intraepithelial carcinoma / Bowen's disease. 13, 14, 34–36, 42, 46–50, 52–54, 56, 59–62, 65, 66

API Application Programming Interface. 41

AUC Area Under the Curve. 9, 11, 15, 24, 43, 47, 68

BCC Basal cell carcinoma. 13, 14, 34–37, 42, 46–50, 52–55, 58–61, 64–67

BKL Benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses). 13, 14, 34–37, 42, 46–49, 51–54, 56, 58, 61, 62, 66, 67

CNN Convolutional Neural Network. 9, 11, 19, 24, 28, 41

DF Dermatofibroma. 13, 14, 34–37, 42, 46–48, 51, 52, 54, 57–59, 61, 63, 66, 67

DL Deep Learning. 9, 28

HAM10000 Human Against Machine with 10000 training images. 13, 34, 46, 68

IA Inteligência Artificial. 11

IG Integrated Gradients. 65–67

INCA Instituto Nacional do Câncer. 23, 24

INFD Infidelity. 15, 65, 66

MEL Melanoma. 13, 14, 34–37, 42, 46–51, 53–55, 57, 59, 61, 64, 66, 67

ML Machine Learning. 9, 26–28, 40

NV Melanocytic nevi. 13, 14, 34–37, 42, 46–49, 51, 52, 54, 56, 59–61, 65–67

RNN Recurrent Neural Networks. 28, 41

ROC Receiver Operating Characteristics. 43

SENS Sensitivity. 15, 65, 66

UV Ultraviolet. 22

UVR Ultraviolet Radiation. 21

VASC Vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage). 13, 14, 34–37, 42, 46–49, 51, 52, 54, 55, 57–59, 61, 63, 66, 67

XAI Explainable Artificial Intelligence. 9, 11, 19, 20, 24–29, 39, 43, 45, 48, 49, 52, 53, 59, 61, 65, 67, 68

Summary

Introduction

According to the World Health Organization (2002), skin cancer has an incidence of 2 or 3 million registered cases worldwide each year. One of the best known potential causes for the disease is sun exposure, which can be enhanced by other factors such as heredity, chronic wounds and scarring. Its main manifestation is the appearance of skin lesions, and the aspects of asymmetry, border, color, dimension and evolution should always be observed: the more asymmetrical, irregular, colored and manifesting growth over time, the greater the chances of being a malignant lesion (Sociedade Brasileira de Dermatologia, 2017).

About sun exposure, the *Global Burden of Disease of Solar Ultraviolet Radiation* (2006) report, showed that the sun's ultraviolet radiation has a great impact on the global emergence of diseases, with malignant melanoma as the most serious consequence. This report estimated that overexposure to Ultraviolet Radiation (UVR) could be responsible for up to 60,000 deaths per year, of which about 48,000 could be classified as malignant melanoma and 12,000 as skin carcinomas. Until that date, it was estimated that the incidence of melanoma had tripled in the then last 45 years in countries such as Norway and Sweden and doubled in the United States over a period of 30 years.

In this sense, for several reasons, human behavior is of great importance when it comes to prevention, from the depletion of the ozone layer (which provides a protective filter) to day-to-day behaviors of sun exposure at specific times, using a filter solar and protective clothing. The *Intersun, WHO's Global UV Project* (2003) report, mentions that measures such as the latter involving the acquisition of new lifestyle habits can extinguish up to 70% of cases of skin cancer in several countries.

The emission of greenhouse gases represents a great concern and threat to the health of the skin, mainly due to the effect that promotes damage to the ozone layer, causing an increase in the passage of ultraviolet rays on the earth's surface and, consequently, greater exposure to human life. This concern led to the enactment of the *Montreal Protocol* (1989), with a report on substances causing the depletion of the ozone layer, which was responsible for the drastic reduction in the concentrations of greenhouse gases (PARKER, 2020).

Also according to Parker (2020), even with commendable improvements in this aspect, it is worth remembering that the long life of substances that deplete the ozone layer have a significantly long life in the atmosphere, which generates the estimate that the total destruction of ozone present in Antarctica does not recover until 2060, generating a steady increase in the release of Ultraviolet (UV) rays for a long period of time, consequently in global rates of skin cancer incidence.

Regarding the promotion of skin health, together with the concern with the sun exposure scenario present in all ages, an initiative of the *Intersun Project* (2003) was to create the so-called *Sun Protection: An Essential Element of Health-Promoting Schools* 2002, which is a package of proposals, guides and practical materials for teachers and schools that aim to promote knowledge of the risks of exposure to sunlight at various levels of primary education, promoting awareness that will impact future generations.

On the other hand, a behavior directly linked to culture and has been shown to be harmful is the use of tanning beds. The *Artificial tanning devices: public health interventions to manage sunbeds* (2017) report showed that, over a period of more than three decades, the deliberate exposure to ultraviolet radiation caused a decrease in the age of onset of skin cancer cases, increasing still its incidence. This fact led the World Health Organization (2005) to recommend that no person under the age of 18 should use tanning beds.

Given the seriousness and advance in the incidence of skin cancer cases, both in Brazil and around the world, there is a need to make the verification of skin spots faster and more accessible. In this sense, the feasibility of this study is classified as seeking to optimize the process of diagnosing skin diseases, allowing for a better distribution of resources, greater agility in the prevention process and, above all, avoiding serious cases of diseases that can lead to death even with a high healing potential, as long as they are identified early (such as the aforementioned malignant melanoma).

1 Background

Whereas one of the main reasons for the dedication of scientists from different areas of knowledge is to know the patterns of diseases, or harmful or positive habits for the existence of human life (and its maintenance), the construction of new technologies and the development of Artificial Intelligence as a field, it becomes a new form of human evolution by making the process of recognizing diseases more efficiently, optimizing time and resources in the identification and treatment of these diseases.

For Kaul, Enslin and Gross (2020) the use of AI in medicine brings numerous benefits, which are cited aspects of an improvement in diagnostic accuracy, as well as a better workflow for professionals and clinics and the general improvement of results also for patients. Despite what the authors called a "lack of general interest" experienced by the AI field in the 70s, 80s and 90s, (result of the knowledge of the limitations in the field and excessive cost to maintain the databases), dates from 1975 the first AI medical workshop, hosted by Rutgers Research Resource on Computers in Biomedicine. For Kulikowski (2015), the moment was favorable for discussions and debates between groups of approaches to AI with similar interests about the field and its potential for medicine.

In the past twenty years, there have been leaps in development in AI. The computer system called "Watson" created by IBM in 2007 in order to be able to answer open-domain questions, won in 2011 the two best players in the game Jeopardy!. According to Ferrucci et al. (2013), the game works with a series of questions from a wide range of information, made for 3 participants who compete against each other, and demands from its participants a quick ability to answer natural language questions, in addition to the ability to interpret (since they often use complex questions with puns and ambiguities), high ability to retain information and linguistic knowledge.

The Watson system developed made use of DeepQA which "is a software architecture for natural language content analysis in questions and knowledge sources" (FERRUCCI et al., 2013), which searches for and analyzes possible valid responses, in addition to punctuating evidence for the same answers. According to Kaul, Enslin and Gross (2020), this technology could be useful to generate evidence-based responses after collecting data from electronic patient data sources.

With regard to the dermatological area, there is a long way that Artificial Intelligence technology can contribute when considering the advancement of skin lesions and diseases such as cancer. In Brazil, skin cancer accounts for 33% of all diagnoses. In addition to this, Instituto Nacional do Câncer (INCA) confirms the increase in incidence by registering, each year, about 180 thousand new cases (Sociedade Brasileira de Dermatologia, 2017). In an estimate for the next 2020-2022 period of INCA, a total of 176,930 cases (93,160 cases in women and 83,770 in men) of non-melanoma skin cancer are expected each year, with a higher incidence in the South, Center West and Southeast of the country.

In the United Kingdom, in the Medicine Journal, Rai (2017) reports in his article four key points: the first explains that benign skin lesions are more common than malignant ones, stating that correctly diagnosing the lesion is fundamental to reassuring patients. In the second key point, Rai (2017) affirms that benign skin lesions can appear suddenly, being that it usually grows or changes gradually or none of these options. In the third key point, the author reports that often benign skin lesions only need treatment if they are symptomatic (for example, pain, itching or social discomfort). In the fourth key point, Rai (2017) states that skin lesions of recent appearance and an unexpected appearance of a skin lesion in patients with a strong family history of malignant skin cancer should be observed with caution.

In view of the need for improvement, it is worth adding how Artificial Intelligence can help in the dermatological context. Bissoto et al. (2019) proposed a set of experiments that reveal some types of positive and negative biases. They concluded that when a model learns to classify malignant lesions by analyzing only the skin (without the edge formations, biological signs or diameter of the lesions), it is strongly dependent on the patterns introduced during the acquisition of the image and the general bias of the dataset.

The work of Esteva et al. (2017), classifies skin cancer at a dermatological level with deep neural networks. There, a CNN was trained using a dataset of 129,450 clinical images consisting of 2,032 different diseases. CNN achieves performance on par with demonstrating an artificial intelligence capable of classifying skin cancer with a competence comparable to dermatologists. The study showed that the classification of skin cancer lesions with artificial intelligence can potentially provide a low-cost form of universal access for an accurate diagnosis.

Mendes and Silva (2018), conducted a skin lesion classification survey using CNN with clinical images. The research resulted in an AUC of 0.96 for melanoma and 0.91 for Basal Cell Carcinoma. This research aims to continue the work of Mendes and Silva (2018), with the application of XAI techniques.

2 Explainable Artificial Intelligence (XAI)

Considering the way we think and how we seek explanations for our behaviors, it is possible to contextualize the real need to understand how the AI's learning process is constituted, since the construction of neural networks for the development of the learning process must be done ethically and as clearly as possible. Within this context, the concept of XAI, emerges as an explainable and responsible need for the construction and use of AI's (KIM, 2018, p. 1).

Explanable Artificial Intelligence (or simply XAI) is conceptualized as a methodology for "large-scale implementation of AI methods in real organizations, with rigor, model explicability and core responsibility" (ARRIETA et al., 2019, p.1).

According to Lin et al. (2019), the purpose of the XAI methodology is to "produce an interpretation for a decision made by a machine learning algorithm" and there is also a specific interest in interpreting as Deep neural networks make decisions, given the complexity and nature of the "black box" of these networks.

As you can see, this is a new field, so little explored. Even so, there is a movement around the scientific field of AI to explore and study performance appraisals in explainability methods. Thus, the XAI represents more than a field of explanation, an area of knowledge of obvious need and opportunity to discover how AIs decide and learn, which can be a milestone in what we know about AI and the ethical development of AIs.

2.1 What is XAI, what is it for, and why use it?

D. Gunning (ARRIETA et al., 2019, p. 6), defined the Explainable Artificial Intelligence as being the creator of a set of machine learning techniques, which in turn, ends up allowing human users to understand, trust in properly and effectively manage the emerging generation of these artificially intelligent partners. Arrieta et al. (2019) stated that this definition brings together two important concepts to be addressed in advance: understanding and trust.

Still according to the author, “the details or reasons that someone gives to make something clear or easy to understand” (ARRIETA et al., 2019, p. 6). In a singular way, the author explains that this definition is a first contribution of the overview provided by him, since the definition assumes that, in different applications, the clarity and ease of understanding of the XAI techniques in the model in question are reversed. For him, explicability remains linked to post-hoc explicability, since it encompasses the techniques used to convert an "uninterpretable into an explicable" model.

What is XAI for?

Now, having understood the concept of XAI, as well as its justification, it is necessary to present its usefulness. Arrieta et al. (2019) state that even a small number of revised articles showing full agreement on the goals needed to establish a description that an explainable model should apply, all of these different goals help differentiate the purpose for which the exercise of the explicability of ML is made, although these few contributions have defined the objectives from a conceptual point of view. Thus, the author summarized some definitions for the aforementioned objectives that XAI has in an attempt to create what he called the "first classification criteria" for the complete collection of articles he addressed in his review. The definitions are described below:

- Trustworthiness

  The term trustworthiness as the main objective of an XAI model is a reason for agreement among several authors. However, claiming that a model is as explicable as its own ways of stimulating confidence may not be what the explicability of the model needs. "Trustworthiness might be considered as the confidence of whether a model will act as intended when facing a given problem."(ARRIETA et al., 2019, p. 7). Thus, the author argues that while trustworthiness is a necessary property of any explainable model, this does not mean that inspiring models of trust in general can be considered explainable on their own, since quantifying trustworthiness is not a easy task. Moreover, this is not the only objective of the explainable model, since the relationship between the two (trust and the explainable model) is not reciprocal. Although there is a consensus among the authors in their articles that they mention the concept of confidence in stating their goal of explicability, they are a minority in light of recent XAI-related research.

- Causality

  This is another of the goals that causality has. The inference of causal relationships based on observational data has been extensively studied over time, and the community working on this topic also widely recognizes it. Thus causality makes it necessary to have a broad background of prior knowledge in order to prove that the effects on observation are indeed causal. "A ML model only discovers correlations among the data it learns from, and therefore might not suffice for unveiling a cause-effect relationship" (ARRIETA et al., 2019). However, the author states that causality involves correlation: an explicable ML model can validate the results provided by causality inference techniques, or it can provide a first guess of possible causal relationships in the available data. But even though causality is a goal of explicability, it is still not among the most important if we look at the number of articles that explicitly state it as their goal.

- Transferability

  According to the author, explicability is also an advocate of transferability, as it has the ability to facilitate the task of clarifying the limitations that may interfere with a model, which allows for better understanding and implementation. Similarly, simply understanding the internal relationships that occur in a model, ultimately facilitates the user's ability to reuse this knowledge in another problem. Moreover, according to Arrieta et al. (2019), "transferability should also fall between the resulting properties of an explainable model, but again, not every transferable model should be considered as explainable".

- Informativeness

  Among one of the precautions that must be taken, being careful not to forget that the problem solved by the model is not the same as a problem faced by its human correspondent is essential. This way, a lot of information is necessary to make possible the relationship between the user's decision and the solution provided by the model and to avoid obstacles and errors. If this is the goal, explainable ML models must be providing information about the problem at hand. Much of the reason found in the articles that were reviewed by the author is "to extract information about the model's internal relationships". He then states that almost all forms of rule extraction serve as the basis for his approach in seeking a simple understanding of what the model internally does, stating that information can be expressed in these simplified proxies, which they believe can explain. the antecedent. This is the argument regularly used in the scientific literature as a way to support what the authors hope to achieve in explicable models.

- Confidence

  Confidence should always be assessed in the model where reliability is expected, with the aim of generalizing firmness and stability. Therefore, the methodology used to keep confidence under control differs depending on the model. Finally, the author considers that "an explainable model should contain information about the confidence of its working regime" (ARRIETA et al., 2019, p. 9).

- Fairness

Socially, explicability can be seen “as the capacity to reach and guarantee fairness in ML models”(ARRIETA et al., 2019). Similarly, the same author adds that support for algorithms and models is increasingly increasing in fields that span human lives, which may mean even greater care with the ethical criteria used. Therefore, explicability should be considered as a filter or bridge that avoids the incorrect, unfair or unethical use of the algorithm outputs.

- Accessibility

  The reaction of unqualified users when dealing with algorithms is often confusing and misunderstood at first glance. In this regard, the author clearly emphasizes that the explicable models will relieve the burden felt by these users and that this concept is expressed in the literature researched by him, as the third most considered objective for the XAI.

- Privacy awareness

  The ability to assess privacy is, according to the author, another possible byproduct of explicability in ML models, noting that they may have complex representations of their learned patterns. Arrieta et al. (2019) points out that “[...] the ability to explain the inner relations of a trained model by non-authorized third parties may also compromise the differential privacy of the data origin” and characterize a violation of it. The author also points out that there is a crucial role that XAI should play in issues such as confidentiality and privacy.

2.2 Deep Learning and CNN explainability

Arrieta et al. (2019) stated that there are two proposed multilayer neural networks that are often used for the explicability studies of DL models: Recurrent Neural Networks (RNN) and CNN. However, given the importance of using CNN and DL in current research, it is at this last point that we will deepen. Moreover, it is worth mentioning some observations in the literature about this.

2.2.1 Challenges to Achieving Explainable Deep Learning

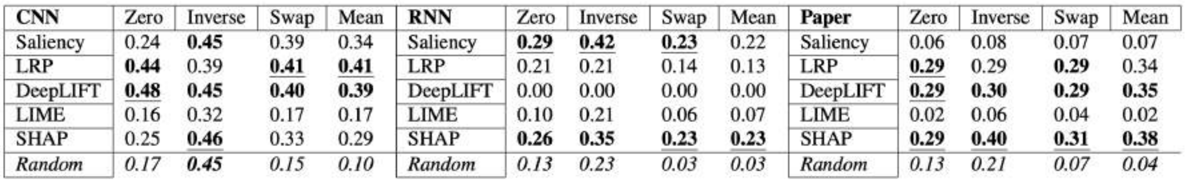
Given the use of Deep Learning for the present work, it becomes relevant to discuss the difficulties that represent a challenge to reach an explainable DL model.

According to Arrieta et al. (2019), there is a lack of agreement about the vocabulary and the different definitions surrounding the XAI, he states that for example, it is constantly possible to see the terms "characteristic of importance" and "characteristic relevance" referring to same concept and this becomes even more evident for visualization methods where there is “absolutely no consistency behind what is known as saliency maps, salient masks, heatmaps, neuron activations, attribution, and other approaches alike” (ARRIETA et al., 2019, p.30).

However, according to the author, he acknowledges that much of this defined absence in the vocabulary is due to the relatively new field of XAI, so the community does not have yet what it calls “standardized terminology”.

2.3 How to apply XAI

Arrieta et al. (2019, p.25) and Mendes and Silva (2018) cited the proposed methods in the LIME method but a study by another author Schlegel et al. (2019) concluded that this LIME method has the worst performance compared to other methods like DeepLIFT and SHAP. Already the methods SHAP and DeepLIFT show a greater robustness as can be seen in the image below:



CNN	Zero	Inverse	Swap	Mean	**RNN**	Zero	Inverse	Swap	Mean	**Paper**	Zero	Inverse	Swap	Mean
Saliency	0.24	**0.45**	0.39	0.34	Saliency	**0.29**	**0.42**	**0.23**	0.22	Saliency	0.06	0.08	0.07	0.07
LRP	**0.44**	0.39	**0.41**	**0.41**	LRP	0.21	0.21	0.14	0.13	LRP	**0.29**	0.29	**0.29**	0.34
DeepLIFT	**0.48**	**0.45**	**0.40**	**0.39**	DeepLIFT	0.00	0.00	0.00	0.00	DeepLIFT	**0.29**	**0.30**	**0.29**	**0.35**
LIME	0.16	0.32	0.17	0.17	LIME	0.10	0.21	0.06	0.07	LIME	0.02	0.06	0.04	0.02
SHAP	0.25	**0.46**	0.33	0.29	SHAP	**0.26**	**0.35**	**0.23**	**0.23**	SHAP	**0.29**	**0.40**	**0.31**	**0.38**
Random	*0.17*	***0.45***	*0.15*	*0.10*	*Random*	*0.13*	*0.23*	*0.03*	*0.03*	*Random*	*0.13*	*0.21*	*0.07*	*0.04*

Figure 1 – Results of the study by Schlegel et al. (2019)

The DeepLIFT and SHAP methods have Python development libraries and GitHub[1,2] repositories and will be detailed in the following topics (2.4, 2.5).

2.4 Ante-hoc and post-hoc

Holzinger et al. (2019) differentiate two types of XAI, one of them being ante-hoc (before the event in question) and the other posthoc (occurring after the event). The ante-hoc XAI method is used in the model structure, therefore, its application is not allowed for all types of model, since its use requires specific models. Choi et al. (2016), exemplify the application of this model with the Reverse Time Attentlon (RETAIN) method, used to simulate medical practice through the attendance of Electronic Health Records (EHR) data in reverse order of time, so that clinical visits made recently are likely to get more attention (CHOI et al., 2016).

On the other hand, the application of the XAI posthoc methodology occurs at the end and can be used for any model, such as the ResNet-152 model, which will be

[1] DeepLIFT repository <https://github.com/kundajelab/deeplift>. Acessed 11/23/2019
[2] SHAP repository <https://github.com/slundberg/shap>. Acessed 11/23/2019

detailed in section 4.3.1. The following section details the posthoc methods that were used in the present study.

2.4.1 DeepLIFT

In the article Learning Important Features Through Propagating Activation Differences (SHRIKUMAR; GREENSIDE; KUNDAJE, 2017), where DeepLIFT (Deep Learning Important FeaTures) is introduced, which, according to Shrikumar, Greenside and Kundaje (2017), is a method used for decomposing the output prediction of a "a specific input by backpropagating the contributions of all neurons in the network to every feature of the input". Your GitHub repository features 490 commits, 20 issues, 3 pull requests, and 8 contributors[1].

2.4.2 SHAP

According to Lundberg and Lee (2017), the SHAP (SHapley Additive exPlanations) method is responsible for identifying a new class of relevant measures on additive characteristics and theoretical results. This demonstrates the existence of a unique solution in this class that has a set of what the author calls "desirable properties". The GitHub SHAP project repository currently has a total of 1048 commits, 399 issues, 17 pull requests, 68 contributors and strong support for code evolution and maintenance[2].

2.4.3 DeepSHAP

Answering the question that asks whether there is a way to make better use of knowledge about the compositional nature of deep networks in order to optimize computational performance, Lundberg and Lee (2017), suggest the answer through a link between the Shapley values and DeepLIFT where DeepLIFT has the function of approximating the SHAP values, assuming a depth of the linear model while also assuming that the input aspects are independent of each other.

2.4.4 GradientSHAP

Like DeepSHAP, GradientSHAP also assumes that the explanation model is linear, as well as the input resources are independent. However, GradientShap approximates the SHAP values by calculating the expectations of gradients by random sampling the layout of baselines. The expected values of gradients are represented by the final values of SHAP (LUNDBERG; LEE, 2017). By means of a white noise inserted to each sample of input `n_samples` times, while selecting a baseline and a random point (along the path between the input and the baseline), calculating the gradient of the outputs in respect

2.4.5 Occlusion

The method proposed by Zeiler and Fergus (2013), deals with a methodology that is based on perturbation in order to calculate attribution, which involves replacing each contiguous rectangular region with a stipulated baseline, computing the difference in the output. When resources are arranged in several regions (hyper rectangles), the output differences are calculated to compute the resource allocation.

2.4.6 GradCam

Proposed by Selvaraju et al. (2016), Gradient-weighted Class Activation Mapping (Grad-Cam) is a technique responsible for producing a series of what the authors called "visual explanations" for the decisions of a wide class of models on CNN. According to Selvaraju et al. (2016), this is an approach that uses gradients of target concepts, generally applied in the last convolutional layer in order to highlight the important regions of the image to predict the concept. GradCam's differentiating factor from other approaches is its applicability to an extensive variety of CNN models, with fully linked layers or CNNs with use for structured exits, for example (SELVARAJU et al., 2016).

2.4.7 Integrateds Gradients

The method combines the "Implementation Invariance of Gradients along with the Sensitivity of techniques like LRP or DeepLift" (SUNDARARAJAN; TALY; YAN, 2017). Unlike other approaches, integrated gradients do not need any network instrumentation, in addition they can be easily computed when using some calls for the gradient operation, which in turn allows even professionals with little experience to be able to apply the technique.

3 Datasets

3.1 Initial selected datasets

Initially were selected 5 different datasets with images of various skin lesions and their respective names. It is noteworthy that the disease names of each image were selected and evaluated by duly qualified professionals from the dermatological area. All selected datasets have been authorized for use in this final project and have all rights reserved. The chosen datasets will be detailed in the sections 3.1.1, 3.1.2, 3.1.3, 3.1.4, 3.1.5 and 3.1.6.

3.1.1 DermNet NZ

DermNet Nz was founded by Dr. Amanda Oakley and has a Health-on-The-Net (HON) certificate[1] since 1996. In addition, it has several awards[2]. According to DermWeb, this dataset has over 20,000 images, about to relaunch with 50,000 images[3].

3.1.2 DermIS

DermIS is a dataset resulting from cooperation between the Dept. of Clinical Social Medicine (Univ. of Heidelberg) and the Dept. of Dermatology (Univ. of Erlangen). According to DermIS, he owns a large collection of links with web pages of hospitals, medical journals and much more[4]. According to DermWeb, this dataset has over 6,800 images[3].

3.1.3 Atlas Dermatológico

The Dermatological Atlas is a Brazilian dataset created by Samuel Freire da Silva[5]. According to the site itself, this dataset currently has 10,409 images, but the dataset is often fed with new images, according to the site[5].

[1] HON certificate <https://www.healthonnet.org/HONcode/Conduct.html?HONConduct455993>

[2] About us <https://www.dermnetnz.org/about-us/>

[3] DermWeb <http://www.dermweb.com/photo_atlas/>

[4] About DermIS <https://www.dermis.net/dermisroot/en/home/index.htm>

[5] Atlas Dermatológico <http://www.atlasdermatologico.com.br>

3.1.4 Edinburgh Dataset

This dataset is provided by the Edinburgh Dermofit Image Library and is publicly available for purchase, under an agreement of a use license[6]. This dataset has 1,300 images.

3.1.5 Dermnet Skin Disease

This dataset was created by Thomas Habif in 1998 in Portsmouth, NH[7]. This dataset has over 23000 images[7].

3.1.6 MED-NODE

This is a small dataset of the Department of Dermatology of the University Medical Center Groningen (UMCG) with only 170 images: 100 naevus images and 70 melanoma images (GIOTIS et al., 2015).

3.2 Final selected dataset

As there are many different datasets and some are updated frequently (such as Atlas Dermatológico and DermIS), standardized datasets were sought to be easily found and used in future work, so the HAM10000 dataset was found and has been chosen to use in this work and will be described below.

3.2.1 HAM10000

The HAM10000 is a large collection of skin lesions images dataset. It has 10,015 images of AKIEC, BCC, BKL, DF, MEL, NV and VASC (TSCHANDL, 2018).

Table 1 – Number of images.

Lesion	HAM10000 images
AKIEC	327
BCC	514
BKL	1,099
DF	115
MEL	1,113
NV	6,705

Continued on next page

[6] Edinburgh Dermofit License <https://licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html>

[7] Dermnet About Us <http://www.dermnet.com/about-us/>

Table 1 – continued from previous page

Lesion	HAM10000 of images
VASC	142
Total	**10,015**

As can be seen, some lesions present many more images than others, in some cases reaching more than 58% difference, such as the difference between the NV and the DF. Below we can see the graph that shows better the discrepancy between one lesion and another.

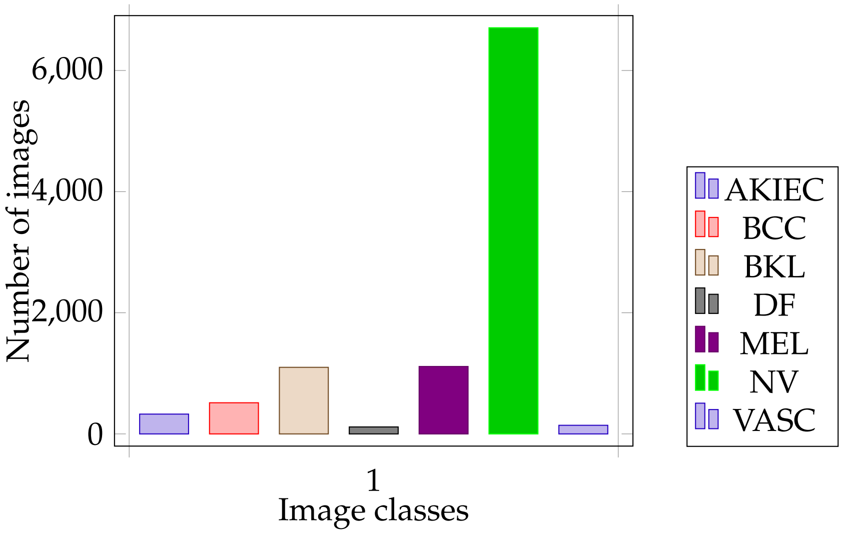


Figure 2 – Image difference graph

In order to get around this problem of this great difference in number of images, a data augmentation was made to expand the dataset and to equalize the number of images between classes which will be explained better in the next section.

3.3 Data Augmentation

Data augmentation is a way to increase the amount of data in a dataset in order to assist in the precision of training a neural network in Deep Learning. In the study by Perez and Wang (2017), the effectiveness of increasing data in the classification of images was tested using SmallNet with 3 different datasets (Dogs vs Goldfish, Dogs vs Cats and MINIST) and it was concluded that the data augmentation has shown to be a promising way to increase the accuracy of image ratings.

In this work, a data augmentation was made to match the number of data for each class of skin lesion images and to significantly increase the number of images in general. The graph below shows the comparison with the number of images in each class.

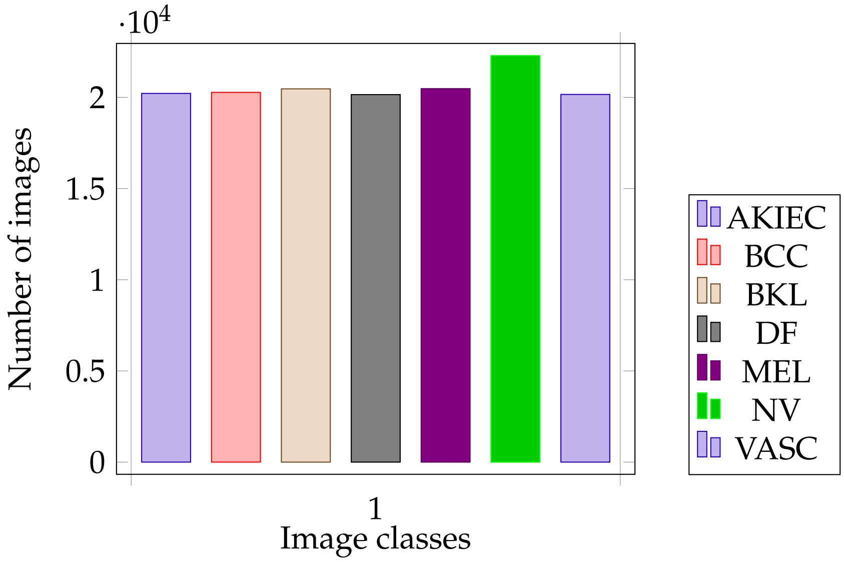


Figure 3 – Image difference with data augmentation graph

The transformations used in the dataset were Rotate, Zoom Random, Flip Random, Flip left-right, Flip top-bottom, Shear, Random Distortion and Lightning, all from the Augmentor library of the article by Bloice, Stocker and Holzinger (2017) which will be further detailed in the Development Libraries and Algorithms. The probability of events for each transformation is detailed below.

Table 2 – Data augmentation probabilities.

Transformation	**Probability**
Rotate	0.5
Zoom Random	0.4
Flip Random	0.5
Flip left-right	0.7
Flip top-bottom	0.5
Shear	0.5
Random Distortion	0.5
Lightning	0.3

In the total of 144,018 images generated by the data augmentation combined with the original images of the dataset, 80% were separated for training and 20% for testing. The table below shows how the separation in the final dataset used in this work was.

Table 3 – Number of images with data augmentation.

Lesion	**Train**	**Validation**
AKIEC	16,139	4,076

Continued on next page

Table 3 – continued from previous page

Lesion	**Train**	**Validation**
BCC	16,166	4,104
BKL	16,259	4,206
DF	16,108	4,041
MEL	16,264	4,212
NV	17,187	5,099
VASC	16,113	4,045
Total	**114,236**	**29,782**

4 Methodology and Metrics

This chapter details the methodology used to perform scientific research and development of parameters for the implementation of XAI in skin lesion image analysis. Section 3.1 deals with the methodology used to construct the theoretical framework. Section 3.1.1 deals with the literature review and section 3.1.2 details the access paths for the data obtained.

4.1 Scientific Research Methodology

A search was made in a database of journals to structure the themes and topics of the theoretical framework, in order to establish a theoretical line for conducting the research. The data should not have a defined publication date, but should theoretically contribute in some way to the construction of the work.

4.1.1 The Review

For the accomplishment of the present work, a systematic literature review was carried out, using concepts of the psychology about human learning to contextualize with the necessity to improve and develop AI learning methods. Bibliographic research has shown that currently there is a small series of publications that bring the importance of XAI (mainly in the year 2019), which demonstrates the growth of productions in the field and an openness to its exploration. Some articles, such as Arrieta et al. (2019), present problems related to the application of the XAI methodology, as well as suggest ways and new possibilities of study.

4.1.2 Selection Criteria

The databases used to search journals were electronic access, such as Scielo, Pepsic, CAPES, Academic Search Ultimate, Scopus and Elsevier. The data must have the proper scientific validity, belonging to recognized institutions, using relevant terms to the contextualization of the theme or to the development of the theme itself. It was not necessary to define a margin of years for the selected scientific productions, due to the use of historical material for a deeper and better understanding of the subject and its current application.

4.2 Initial Development Methodology

For the development had been carried out a planning of which approach and which tools will be used during the evolution of the project. The following sections will show which development support tools, libraries and development algorithms will be used.

4.2.1 Development Support Tools

- GitHub

  GitHub is a platform that uses Git, an open source program that tracks changes to files and folders[1]. This work will take advantage of the optimization that the platform seeks to perform in relation to the described operation providing advanced functionality and easier access through the interaction in graphical interface, since its support can be done for teams or individual projects[1].

- Docker

  According to the documentation[2], Docker is an open platform for application development, submission and execution. It allows applications to be separated from infrastructure to facilitate software delivery, significantly reducing the delay between writing code and running it in production.

- Jupyter Notebook

  Jupyter Notebook is a data science tool that helps you visualize inputs and outputs between lines of code. According to Jupyter Notebook documentation, it uses a console-based approach to interactive computing in a qualitatively new direction, providing a web application suitable for capturing the entire computing process: development, documentation and code execution, and communication of results[3].

4.2.2 Development Libraries and Algorithms

- SHAP

  It is characterized as a unified explanation approach to the output of any ML model. SHAP is responsible for connecting game theory and local explanations, in addition to representing the only consistent additive resource allocation method and precisely possible location based on expectations[2].

[1] GitHub Glossary <https://help.github.com/en/github/getting-started-with-github/>

[2] Docker documentation <https://docs.docker.com/engine/docker-overview/>

[3] Jupyter Notebook Documentation <https://jupyter-notebook.readthedocs.io/en/latest/notebook.html>

- Keras

  According to information from the site[4], the platform is "a high level neural network Application Programming Interface (API) written in Python and capable of running on TensorFlow, CNTK or Theano". It has been designed to enable rapid experimentation as well as being able to convey the idea to the result efficiently (with minimal delays). It allows an easy and agile propagation and supports CNN and RNN as well as combinations of the two.

- Numpy

  Numpy is a basic Python library. According to the site, it contains an N-dimensional array object, sophisticated transmission-related functions, and C/C++ and Fortran integration tools. It is easy to use as it is licensed under the BSD license allowing reuse with few restrictions. In addition it can still be used as a generic data container[5].

- Scikit-learning

  With a Numpy-based system, the Scikit-learning library adds the incorporation of algorithms for phyton ML learning tasks, as well as analyzing and mining data. According to information from the site itself, the library is accessible and reusable in a variety of contexts, as well as BSD licensed and open source[6].

- Caffe

  Caffe is a DL framework developed by Berkeley AI Research (BAIR) and community contributors. Caffe is licensed under BSD 2 - Clause. According to the site information, it can contribute expressive architecture that encourages application and innovation by switching between CPU and GPU; with its extensible code that can promote active development and its speed, which promotes greater efficiency in image processing[7].

- OpenCV

  The Open Source Computer Vision Library (OpenCV) is a free, open source, computer vision and cross-platform library for commercial and academic use, created by Intel in the mid-2000s. Its main goal is a simple infrastructure that enables creations. and sophisticated applications quickly by developers and helpers. It offers a range of computer vision algorithms, such as image filtering, face recog-

[4] Keras <https://keras.io/>

[5] Numpy <https://numpy.org/>

[6] Scikit-learning <https://scikit-learn.org/stable/>

[7] Caffe <http://caffe.berkeleyvision.org/

nition, and object recognition, as well as supporting a variety of languages such as Python, C ++, Ruby, and Matlab[8].

4.3 Final Development Methodology

4.3.1 Neural Network Architecture and fine tuning

As this work is a continuation of the study by Mendes and Silva (2018), which serves as a comparison with the study by Han, the same neural network architecture called ResNet-152 was used. As described by Mendes and Silva (2018) and its original study He et al. (2015), ResNet is a deep neural network architecture that has 152 layers with bottleneck design.

In this work the transfer learning method was done with the ResNet-152 pre-trained with ImageNet dataset as in the work of Mendes and Silva (2018), however with the fine tuning of the layer fully connected applying a linear transformation to the incoming data layers for output of the 7 classes: AKIEC, BCC, BKL, DF, MEL, NV and VASC.

4.3.2 Development Support Tools

- Kaggle

  Kaggle is a platform with online virtual machines from the subsidiary Google LLC that allows the creation of notebooks with the aid of CPU, TPU or GPU. This platform is widely used for machine learning competitions. In this work it was used to assist in obtaining the data with the help of its hardware (since it provides limited hours per week for the use of the Nvidia Tesla K80 GPU)[9].

4.3.3 Development Libraries and Algorithms

- PyTorch

  Python library is based on another library called Torch. According to (PASZKE et al., 2019) is a library that debugs easily, is highly efficient and has support for hardware acceleration such as multiple GPUs. It was used in this study to use the Captum library (mentioned in the section 4.3.3).

- Augmentor

[8] OpenCV <https://opencv.org/>

[9] Kaggle <https://www.kaggle.com/>

it is a library proposed in the article Bloice, Stocker and Holzinger (2017) that helps to make the data augmentation. This library was used in the project to apply methods such as flip, zoom, distortion, rotation, sher and lightning

- Captum

  The Captum library was developed by Facebook engineers Kokhlikyan et al. (2020) in order to assist and unify XAI methods for the PyTorch library. This project is open-souce and is available on GitHub[10]. It was used in this work to implement XAI techniques and to collect the metrics used.

4.4 Metrics

4.4.1 Evaluating neural networks

In the article of Fawcett (2006), was proposed the Receiver Operating Characteristics (ROC) graph introdution. ROC graphs are commonly used in medical decision making, which leads doctors to decide whether or not this skin lesion is a melanoma (SWETS, 1986). This graph provides a value metric called of AUC, a measure of the discriminability of a pair of classes. Sensitivity and specificity are used to validate the diagnosis. The higher the sensitivity (closer to 1.0) and the lower the specificity (closer to 0.0), the better the result and the more valid the diagnosis is (KUMAR; INDRAYAN, 2011; HAJIAN-TILAKI, 2013).

4.4.2 Evaluating XAI

In the study by Ancona et al. (2017) that tries to create a better understanding of gradient-based attribution methods for deep neural networks, a new method was proposed to evaluate metrics for comparing methods for XAI. The Yeh et al. (2019) article, based on the Ancona et al. (2017) article, proposed two new methods for comparing XAI called Infidelity and Sensitivity.

The explanation's infidelity represents the expected mean-square error between the explanation of XAI multiplied by a significant input perturbation and the differences between the predictor function at its input and the disturbed input; the explanation's sensitivity measures the extent of the change in explanation of XAI when the input is slightly pertubed. If the models have a high sensitivity of explanation that have been shown to be prone to adversarial attacks then the interpretation of neural networks is fragile (YEH et al., 2019).

[10] Captum <https://github.com/>

5 Results

The tests were performed in the Kaggle environment using the limited time of the Nvidia Tesla K80 GPU, as mentioned in the section 4.3.2. Version 0.3.1 of the Captum[1] library (section 4.3.3), PyTorch version 1.7.0 (section 4.3.3), ResNet-152 pretrained with ImageNet[2] (section 4.3.1) was used.

In the experiment, the pre-trained ResNet-152 model was trained, it was early stopped by training accuracy with 88% and with an accuracy of 86% for validation, which was defined to stop after 3 epochs without improvement in accuracy or loss of both the training and validation. The SGD optimizer was used with the same parameters as the work of Mendes and Silva (2018) with *learning_rate* 0.01, *momentum* 0.9 and *weight_decay* 1e-05. For the scheduler, the exponential learning rate with *gamma* 0.1 was used and for the loss the cross entropy was used and in the work of Mendes and Silva (2018).

In table 4 below, the total number of trainable and non-trainable parameters will be available, along with size (MB) of inputs and parameters.

Table 4 – Model parameters.

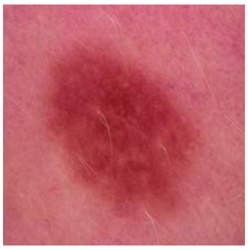
	Parameters
Trainable	58,148,615
Non-trainable	9,536
Total of parameters	**58,158,151**
Inputs size (MB)	0.57
Forward/backward pass size (MB)	606.58
Parameters size (MB)	221.86
Estimated Total Size (MB)	**829.01**

The images chosen for the experiments performed in this work were the figures with more than 80% prediction of each classification of the HAM10000 dataset. The figures will be shown below in the images 4, 5, 6, 7, 8, 9 and 10 with their respective predictions

[1] XAI tests (version 11) <https://www.kaggle.com/matheusherique/xai-captum>.

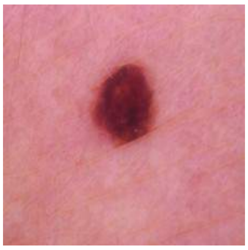
[2] ResNet-152 train (version 17) <https://www.kaggle.com/matheusherique/tcc-pytorch>.

Figure 4 – Image ISIC_0024465 from HAM10000 dataset.



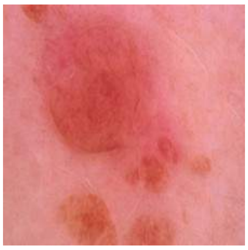
0.852% for NV class.

Figure 5 – Image ISIC_0025423 from HAM10000 dataset.



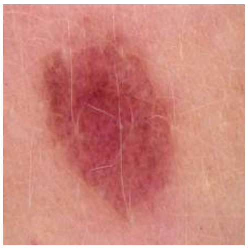
0.997% for AKIEC class.

Figure 6 – Image ISIC_0026064 from HAM10000 dataset.



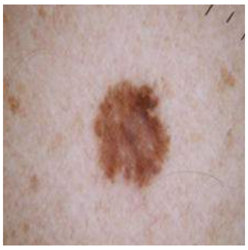
0.974% for BCC class .

Figure 7 – Image ISIC_0024788 from HAM10000 dataset.



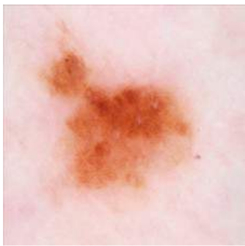
0.972% for BKL class.

Figure 8 – Image ISIC_0024447 generated from HAM10000 data augmentation.



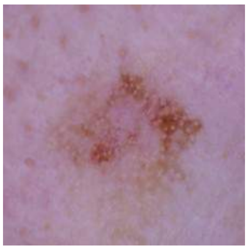
1.000% for DF class.

Figure 9 – Image ISIC_0027461 from HAM10000 dataset.



0.752% for MEL class.

Figure 10 – Image ISIC_0030005 from HAM10000 dataset.



1.000% for VASC class.

The table 5 shows the results of the AUC metrics in this work. These results will be compared in the following table (table 6) with previous studies of Esteva et al. (2017), Han et al. (2018) and Mendes and Silva (2018).

Table 5 – AUC of this work.

Lesion	**AUC**
AKIEC	1.00
BCC	0.99
BKL	0.93
DF	1.00
MEL	0.90
NV	0.79
VASC	1.00

We can see the comparison in the table 6, as results of previous works including this work that is being continued from Mendes and Silva (2018). As previously mentioned in the section 4.4.1, the closer to 1.0 the AUC value the more correct the prediction can be the opposite is also true. The closer to 0.0 is the more wrong the prediction can be.

This work had five classifications that were superior to previous studies, the classification of AKIEC, BCC, BKL, DF and VASC, whereas the classifications of NV and MEL were superior in the studies of Mendes and Silva (2018) and Esteva et al. (2017).

Table 6 – Presenting new AUC results. Unlike the other studies mentioned, the HAM10000 dataset was used in this study.

Lesion	**Esteva et al. (2017)**	**Han et al. (2018)**	**Mendes and Silva (2018)**	**This work**
AKIEC	-	0.83	0.96	**1.00**
BCC	-	0.90	0.91	**0.99**
BKL	-	0.89[3]	0.90[3]	**0.93**
DF	-	0.90	0.90	**1.00**
MEL	0.96	0.88	**0.96**	0.90
NV	-	0.94	**0.95**	0.79
VASC	-	0.83[4]	0.99[4]	**1.00**

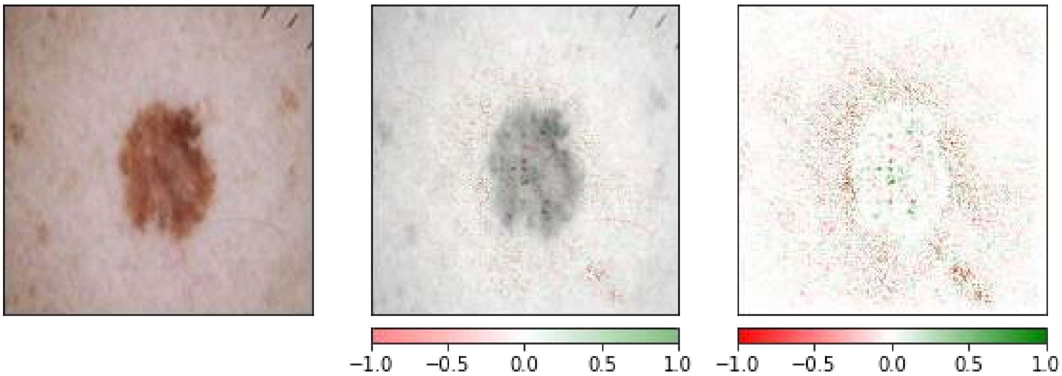
[3] Seborrheic keratosis <https://www.icd10data.com>

[4] Hemangioma <https://www.icd10data.com>

5.1 DeepLIFT

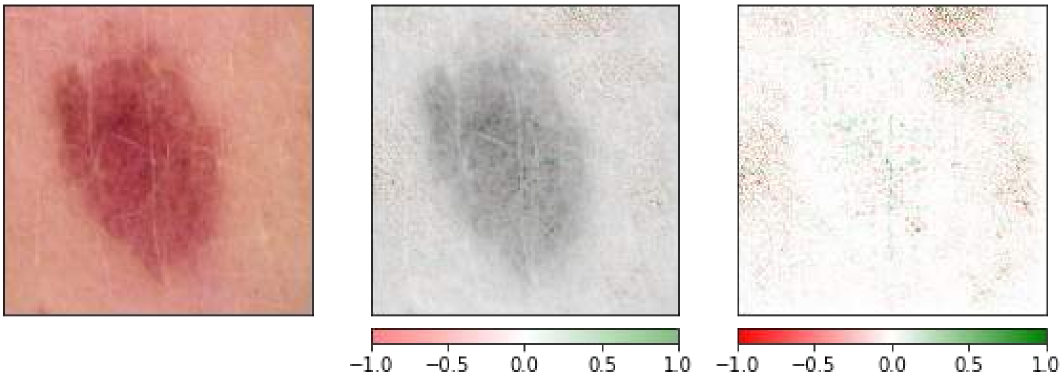
The first tests were performed using the XAI DeepLIFT method. As will be shown later in the tables 8 and 7, this method obtained great results in terms of runtime, Sensitivity and Infidelity. As we can see in the images 11, 12, 13, 14 of DF, BKL, VASC, NV, we can see on the green dots that method was able to visualize the lesions well and ignored the unimportant on the red dots in the skin and hair.

Figure 11 – DeepLIFT applied to DF class.



The DeepLIFT method visualizing DF lesion. Left-to-right: original image, blended heat map image, heat map image.

Figure 12 – DeepLIFT applied to BKL class.

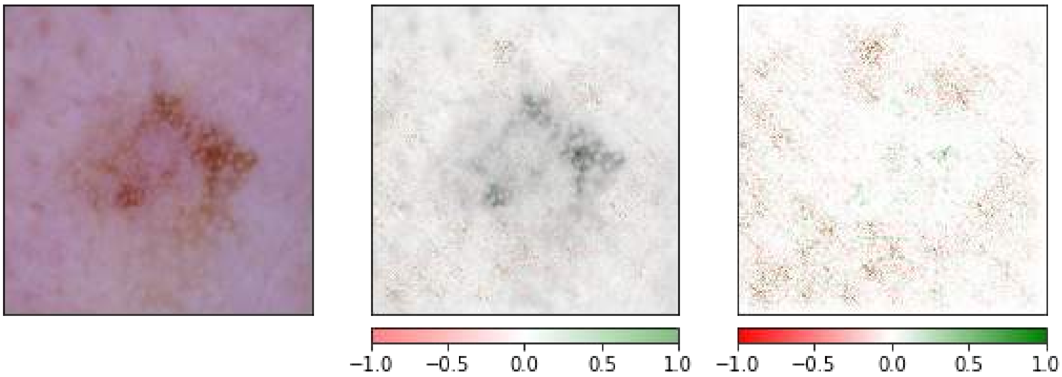


The DeepLIFT method visualizing BKL lesion.

In the BCC (figure 15) and MEL (figure 16) images, the model focused more on less important dots such as the skin and completely ignored the lesions, due to a possible overfitting.

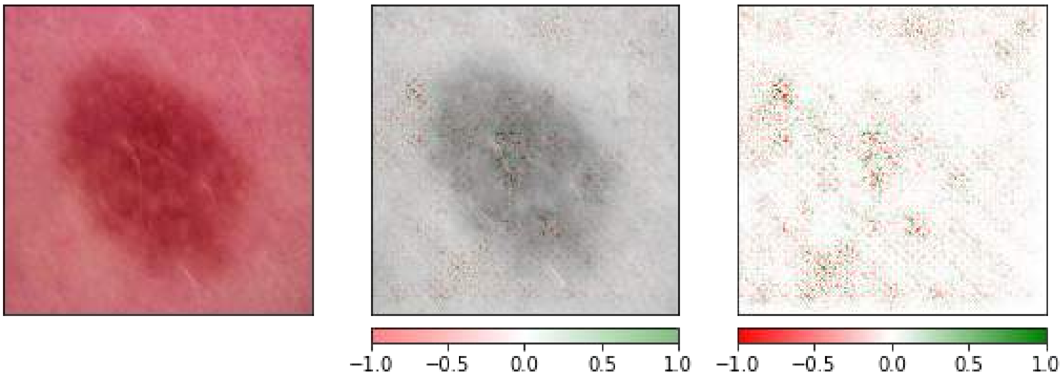
In the AKIEC image, these was a little difficult to identify where the trained model is viewing, as the dots are very dispersed, but with most of the green dots being more focused on the lesion with a border of ignored dots around the lesion.

Figure 13 – DeepLIFT applied to VASC class.



The DeepLIFT method visualizing VASC lesion.

Figure 14 – DeepLIFT applied to NV class.



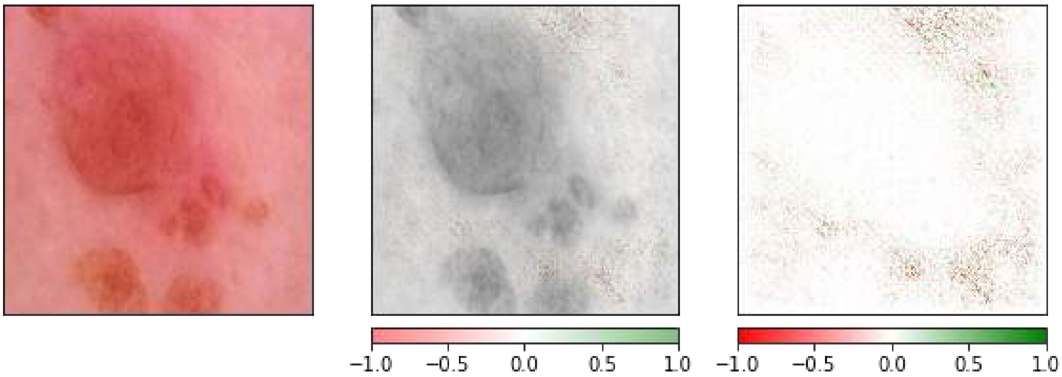
The DeepLIFT method visualizing NV lesion.

5.2 Occlusion

In the tests of the XAI Occlusion model, as will be shown in the tables 8 and 7, they had the worst results for execution time and Infidelity and the best results for Sensitivity, so it may be easier to see where the model is viewing in the image, compared to the other methods tested. As can be seen in the images of BKL (figure 18), MEL (figure 19), NV (figure 20) and VASC (figure 21), they are the easiest figures to be understood where the model is seeing correctly, the green projections are being located on the lesions and the red ones on the skin. An observation for the VASC and MEL classes image, we can see that the green projections are being dispersed to places outside the lesion, considering a good part of the skin as a lesion, which can be explained with the Infidelity metric. These images have greater infidelity (5.602 and 2.185 respectively) between these four classes.

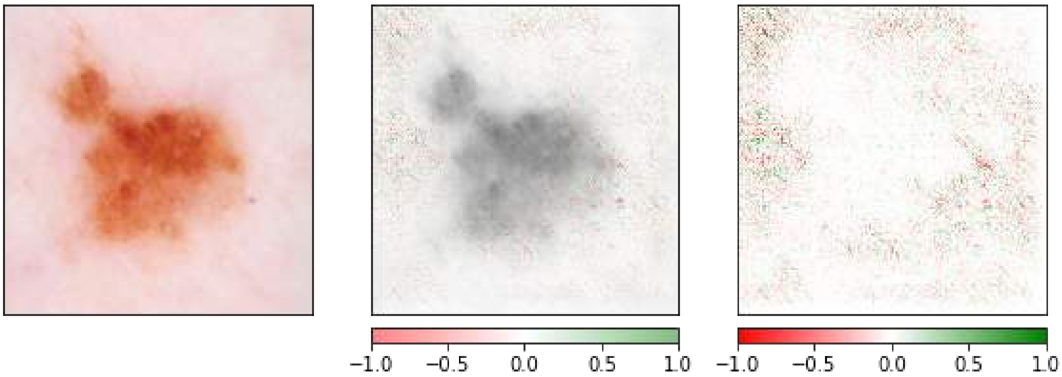
The two images that had the highest Infidelity were the AKIEC (figure 22) and BCC (figure 23) classes. In these images we can see that they had these results for not

Figure 15 – DeepLIFT applied to BCC class.



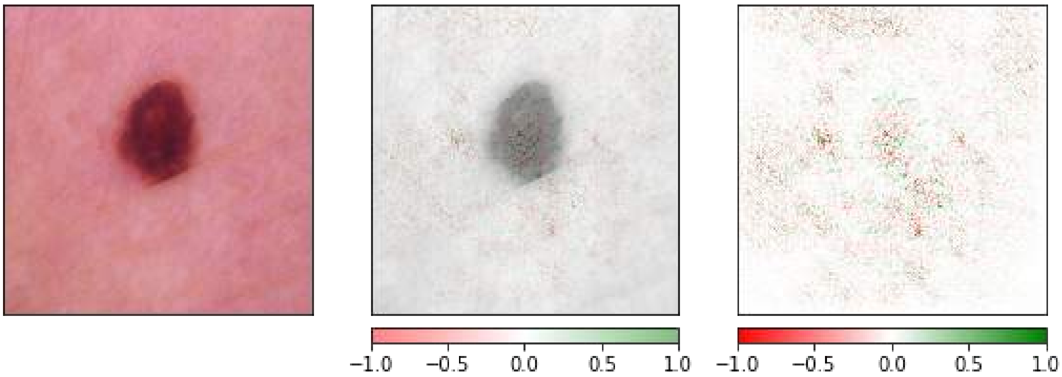
The DeepLIFT method visualizing BCC skin with possible overfitting.

Figure 16 – DeepLIFT applied to MEL class.



The DeepLIFT method visualizing MEL skin with possible overfitting.

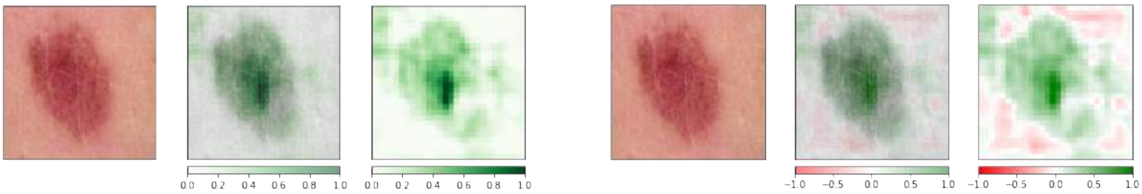
Figure 17 – DeepLIFT applied to AKIEC class.



The DeepLIFT method visualizing AKIEC lesion with dots very dispersed.

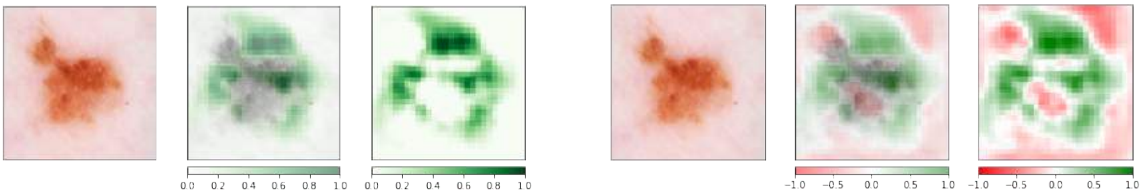
matching the explanation with the image as if he had not focused on anything, neither the lesion nor the skin, that’s why such a high infidelity.

Figure 18 – Occlusion applied to BKL class.



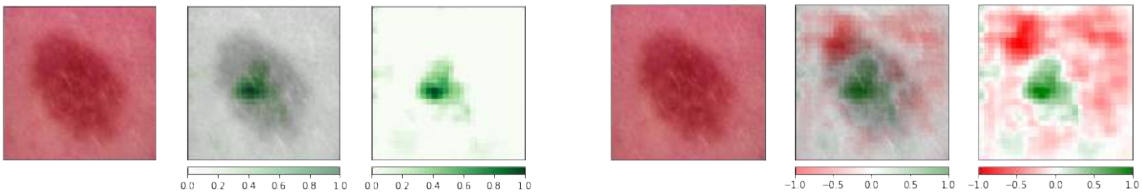
The Occlusion method visualizing BKL lesion. Left-to-right: original image, blended heat map image with positives values, heat map image with positives values; original image, blended heat map image with all values (positives and negatives), heat map image with all values as well.

Figure 19 – Occlusion applied to MEL class.



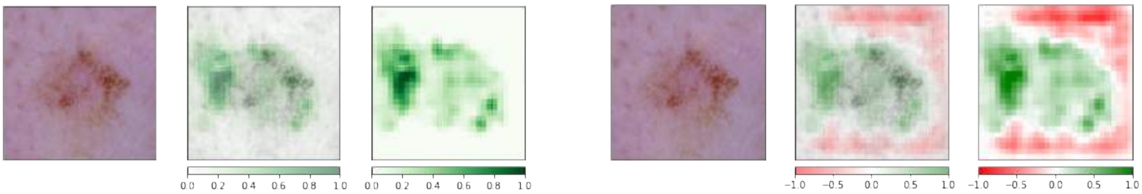
The Occlusion method visualizing MEL lesion.

Figure 20 – Occlusion applied to NV class.



The Occlusion method visualizing NV lesion.

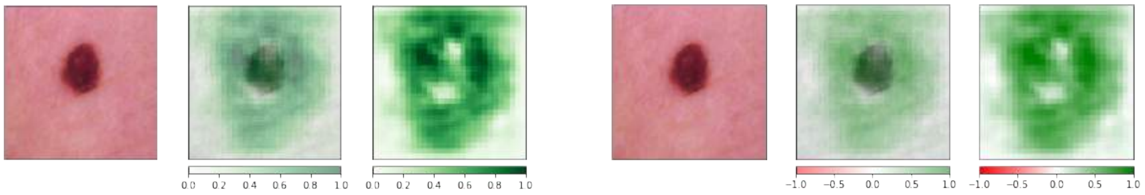
Figure 21 – Occlusion applied to VASC class.



The Occlusion method visualizing VASC lesion.

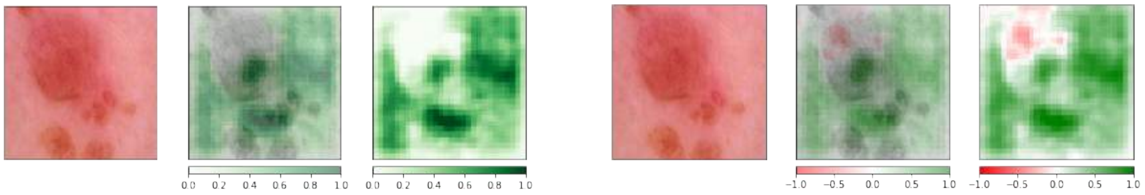
In the DF (figure 24) class, Infidelity is low, but it shows that the model visualize wrong places in the image. Infidelity is low because the points visualized by the model correspond to the image, differentiate the skin from the lesion and separate them with an border. The points seen by the model are wrong because it identified the skin as

Figure 22 – Occlusion applied to AKIEC class.



The Occlusion method visualizing AKIEC lesion.

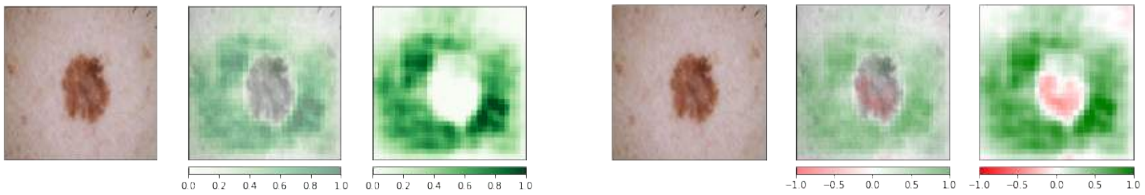
Figure 23 – Occlusion applied to BCC class.



The Occlusion method visualizing BCC lesion.

a positive point and the lesion as a negative point, which was supposed to be the opposite.

Figure 24 – Occlusion applied to DF class.

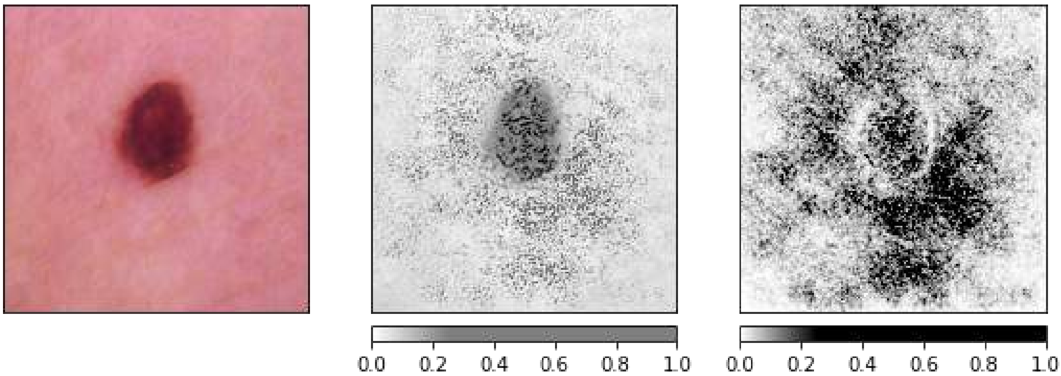


The Occlusion method visualizing DF lesion wrong.

5.3 Integrated Gradients

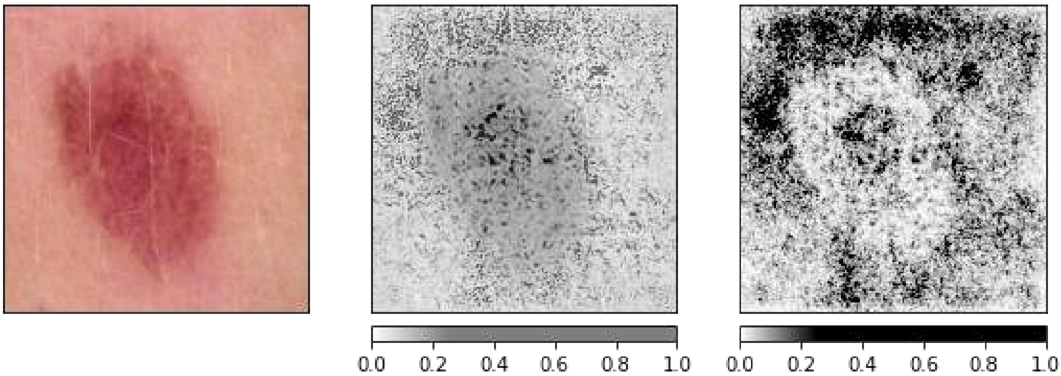
Integrated Gradients works differently from previous methods in this method visualizes points that are more spread out because of gradients. This method will analyze the comparison of Integrated Gradients (with positive values) with Gradients (with positive and absolute values). Integrated Gradients was the XAI method that presented the best Infidelity results, with the results closest to 0. The classes AKIEC (figure 25), BKL (figure 26), DF (figure 27), NV (figure 28) and VASC (figure 29) presented the best positive values, concentrating most of the points in the lesion; the absolute value of these images is focused more on the edge of the lesions and on the skin.

Figure 25 – Integrated Gradients applied to AKIEC class.



The Integrated Gradients method visualizing AKIEC lesion. Left-to-right: original image, positives values, positive with absolute values.

Figure 26 – Integrated Gradients applied to BKL class.



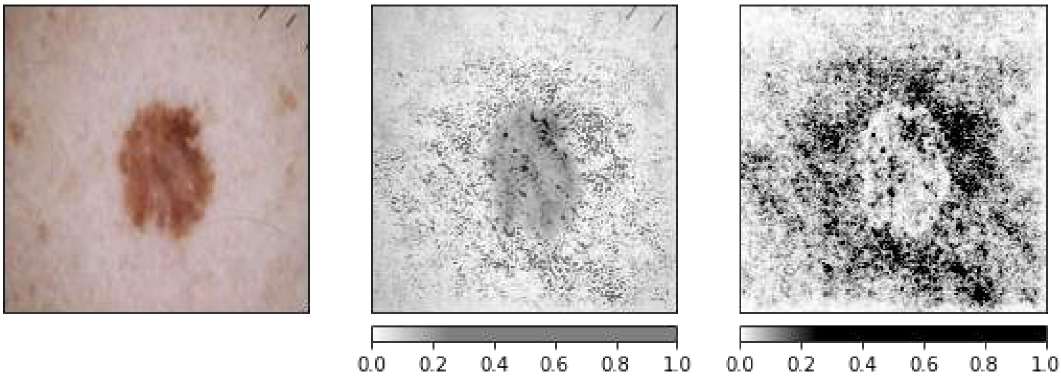
The Integrated Gradients method visualizing BKL lesion. In this image, the positive values were concentrated in a small portion of the lesion, but even so the points were further inside the lesion. Left-to-right: original image, positives values, positive with absolute values.

The BCC (figure 30) and MEL (figure 31) classes did not present many positive values in the inner parts of the lesion, in the image with the absolute values it can be seen that the model concentrated more on the skin surroundings.

5.4 GradientSHAP

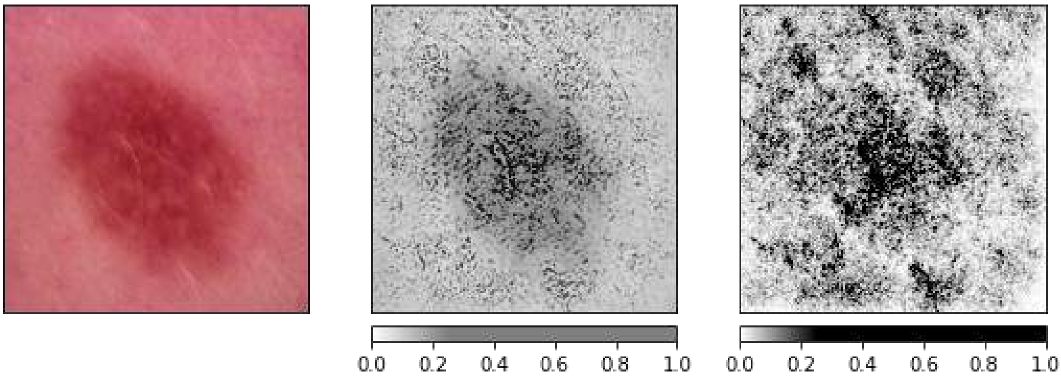
GradientSHAP was the XAI method that had the highest Sensitivity, which measures the degree to which the explanation is affected by insignificant perturbations in the test images. Highly sensitive explanations may be more susceptible to adversarial attacks, which is why GradientSHAP had the worst results in this regard. An observation for this method is that, since it is based on SHAP, each time the code

Figure 27 – Integrated Gradients applied to DF class.



The Integrated Gradients method visualizing DF lesion. In this image, like the previous ones, the positive values were concentrated in a small portion of the lesion, but even so the points were further inside the lesion. Left-to-right: original image, positives values, positive with absolute values.

Figure 28 – Integrated Gradients applied to NV class.



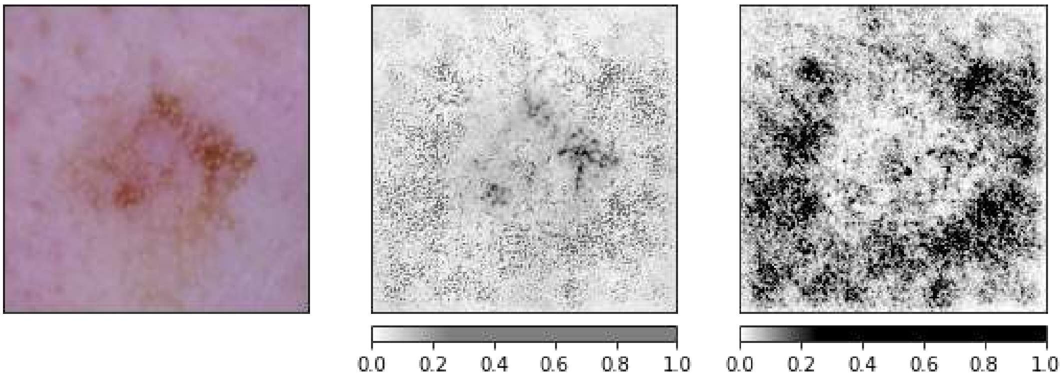
The Integrated Gradients method visualizing NV lesion. Left-to-right: original image, positives values, positive with absolute values.

runs, the method presents gradients in different places in the image because it suffers oscillations.

In this specific test, only the NV class figure (32) was able to focus entirely on the lesion. In the AKIEC (figure 33), BKL (figure 34), DF (figure 35) and VASC (figure 36) images, the image had a mixture of views between the lesion and the skin.

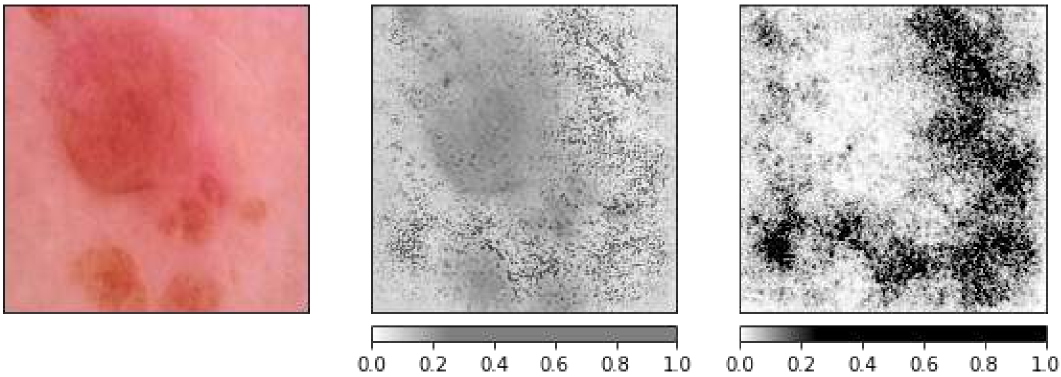
The images of the MEL and BCC classifications showed, in this specific test, dots only on the skin parts, probably because the training had overfitting and learned to differentiate the lesions through the skin around the lesion.

Figure 29 – Integrated Gradients applied to VASC class.



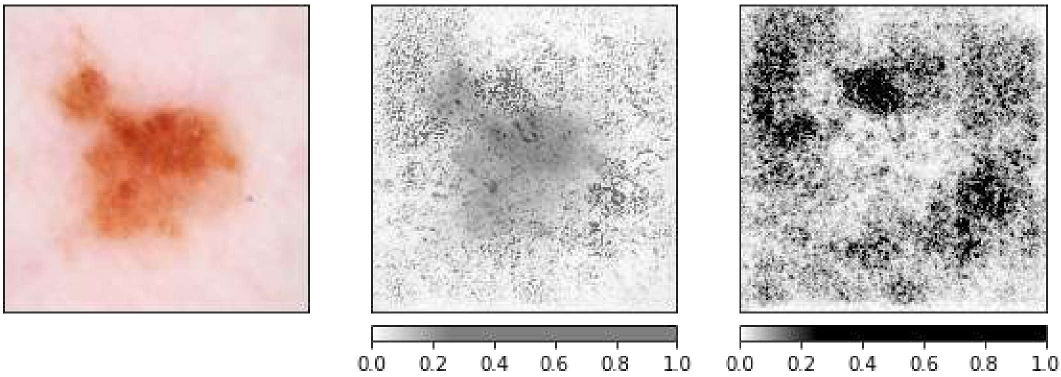
The Integrated Gradients method visualizing VASC lesion. Left-to-right: original image, positives values, positive with absolute values.

Figure 30 – Integrated Gradients applied to BCC class.



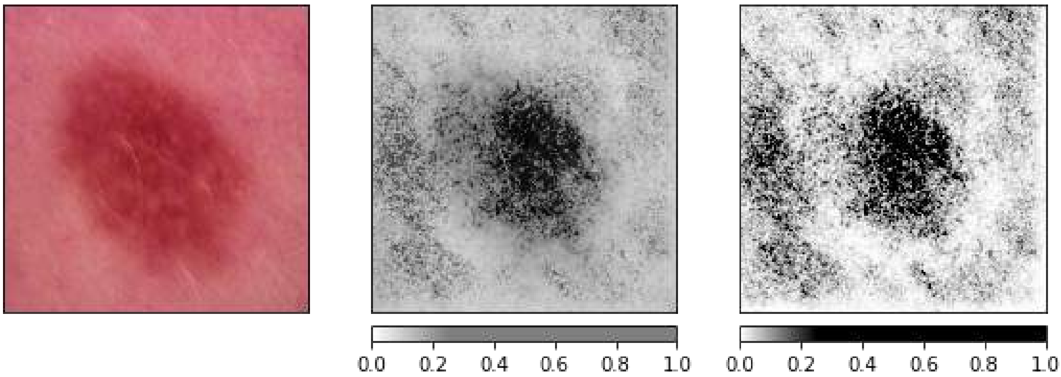
The Integrated Gradients method does not visualizing BCC lesion. Left-to-right: original image, positives values, positive with absolute values.

Figure 31 – Integrated Gradients applied to MEL class.



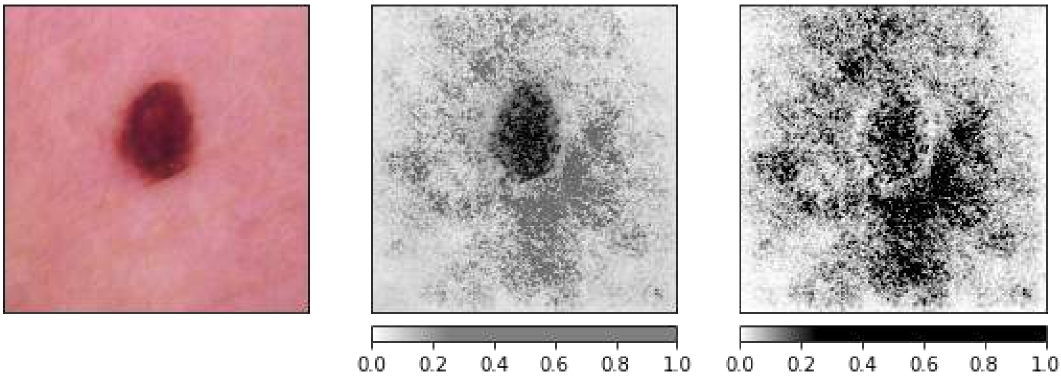
The Integrated Gradients method does not visualizing MEL lesion. Left-to-right: original image, positives values, positive with absolute values.

Figure 32 – GradientSHAP applied to NV class.



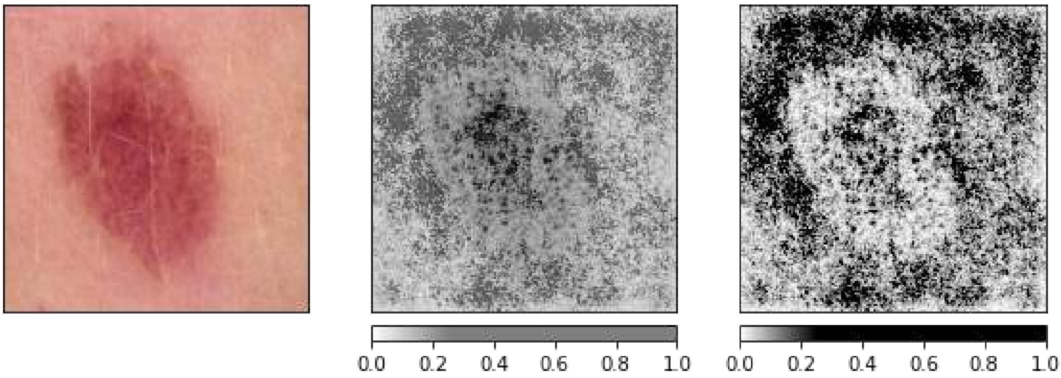
The GradientSHAP method visualizing NV lesion. Left-to-right: original image, blended heat map with absolute values, heat map with absolute values.

Figure 33 – GradientSHAP applied to AKIEC class.



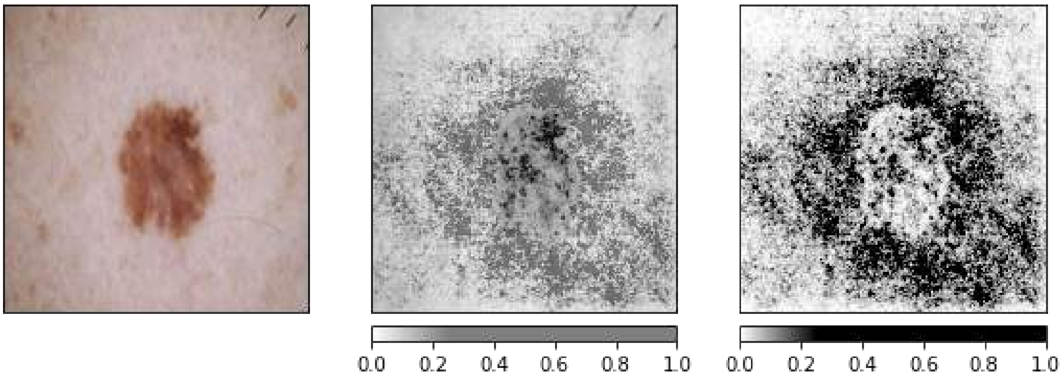
The GradientSHAP method visualizing AKIEC lesion and skin. Left-to-right: original image, blended heat map with absolute values, heat map with absolute values.

Figure 34 – GradientSHAP applied to BKL class.



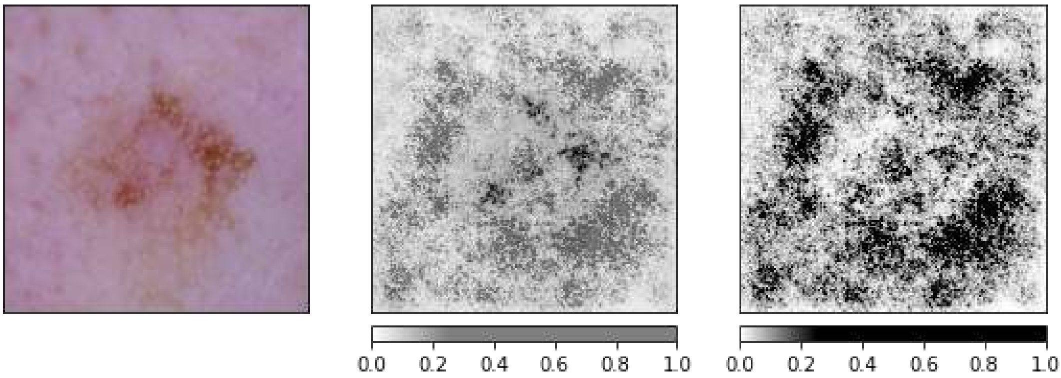
The GradientSHAP method visualizing BKL lesion and skin. Left-to-right: original image, blended heat map with absolute values, heat map with absolute values.

Figure 35 – GradientSHAP applied to DF class.



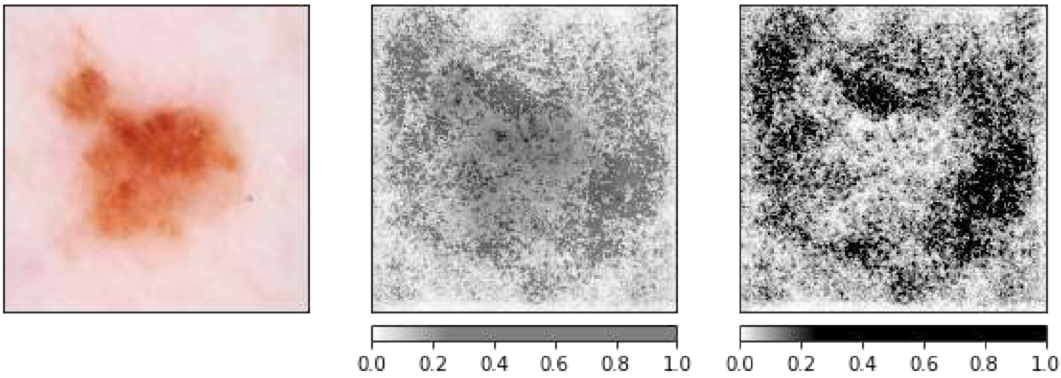
The GradientSHAP method visualizing DF lesion and skin. Left-to-right: original image, blended heat map with absolute values, heat map with absolute values.

Figure 36 – GradientSHAP applied to VASC class.



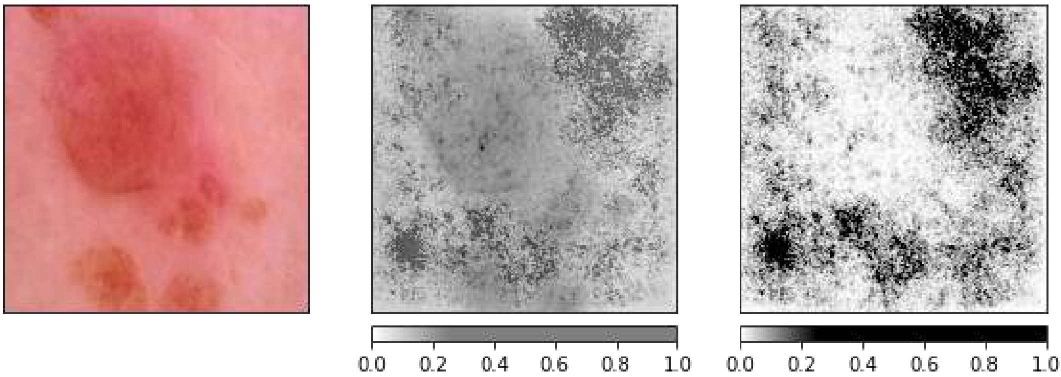
The GradientSHAP method visualizing VASC lesion and skin. Left-to-right: original image, blended heat map with absolute values, heat map with absolute values.

Figure 37 – GradientSHAP applied to MEL class.



The GradientSHAP method visualizing MEL skin. Left-to-right: original image, blended heat map with absolute values, heat map with absolute values.

Figure 38 – GradientSHAP applied to BCC class.



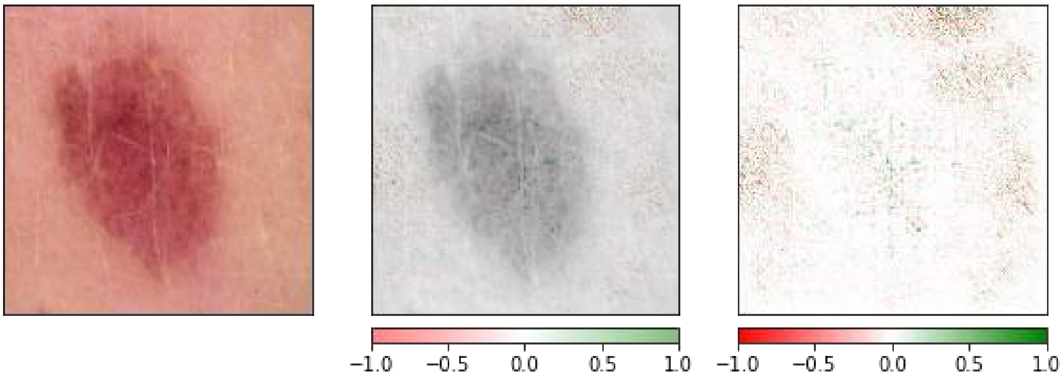
The GradientSHAP method visualizing BCC skin. Left-to-right: original image, blended heat map with absolute values, heat map with absolute values.

5.5 DeepSHAP

As it is a mixture of DeepLIFT and SHAP, two methods that use a lot of the graphics card's VRAM, it was not possible to collect the Sensitivity data by the DeepSHAP method, as there is not enough space to generate a perturbation to images even when clearing the cache in the PyTorch before code execution.

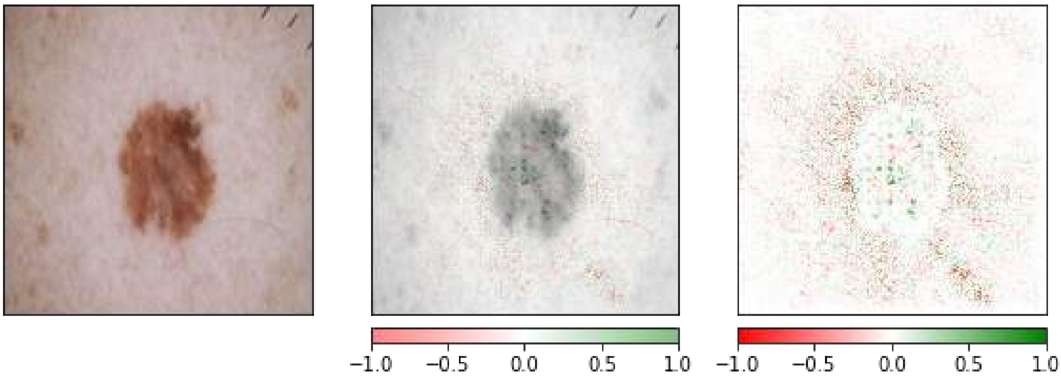
Although it was not possible to obtain information about Sensitivity , this method had one of the best results for Infidelity. The results of this method for the BKL (figure 39), DF (figure 40) and VASC (figure 41) classes were very positive, as it separates the positive values for the lesions and the negative ones for the skin. An observation for the BKL class is that DeepSHAP is seeing more hairs inside the lesion, which will also happen in the next GuidedGradCam method.

Figure 39 – DeepSHAP applied to BKL class.



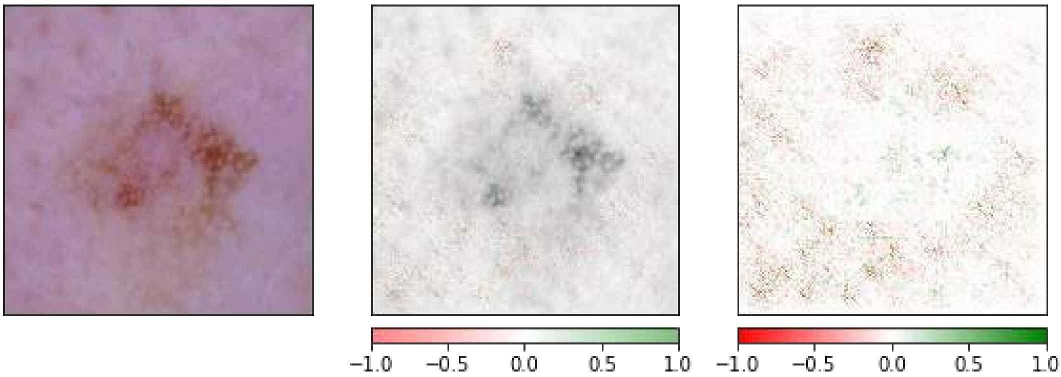
The DeepSHAP method visualizing BKL lesion. Left-to-right: original image, blended heat map with all values (positives and negatives), heat map with all values.

Figure 40 – DeepSHAP applied to DF class.



The DeepSHAP method visualizing DF lesion. Left-to-right: original image, blended heat map with all values (positives and negatives), heat map with all values.

Figure 41 – DeepSHAP applied to VASC class.

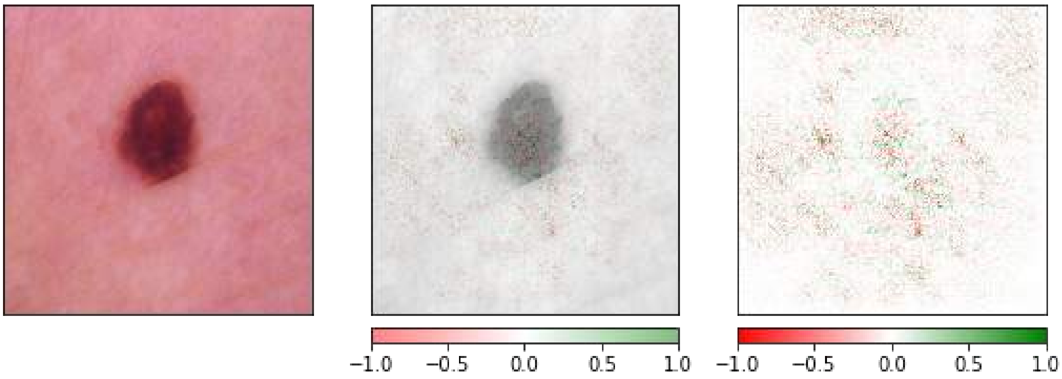


The DeepSHAP method visualizing VASC lesion. Left-to-right: original image, blended heat map with all values (positives and negatives), heat map with all values.

The AKIEC (figure 42) and NV (figure 43) classes the positive and negative points were very dispersed; in the AKIEC class you can see a slight border between the skin and the lesion, there was a lot of positive point in the lesion, but it surpassed the skin in some moments, which was the same case in the NV class.

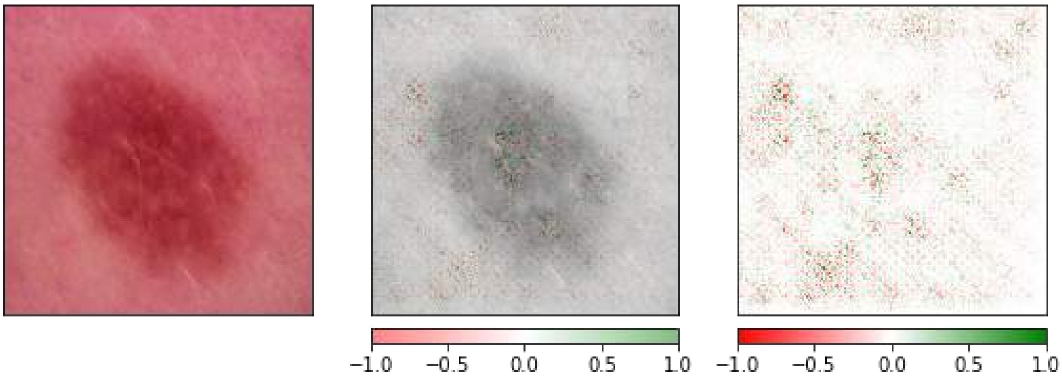
As for the BCC (figure 44) and MEL (figure 45) classes, only the edges of the lesions and the skin were recognized, both positive and negative points, the lesion being almost totally ignored, which could have happened due to possible overfitting, as mentioned in the previous XAI techniques.

Figure 42 – DeepSHAP applied to AKIEC class.



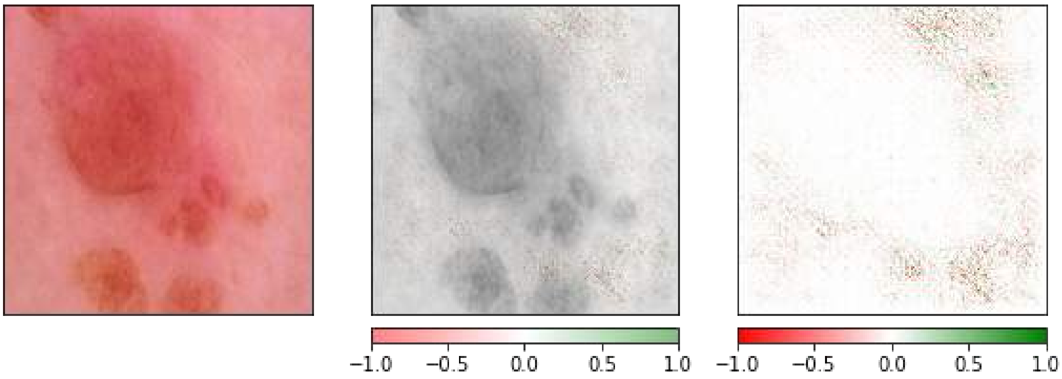
The DeepSHAP method visualizing AKIEC dispersed dots. Left-to-right: original image, blended heat map with all values (positives and negatives), heat map with all values.

Figure 43 – DeepSHAP applied to NV class.



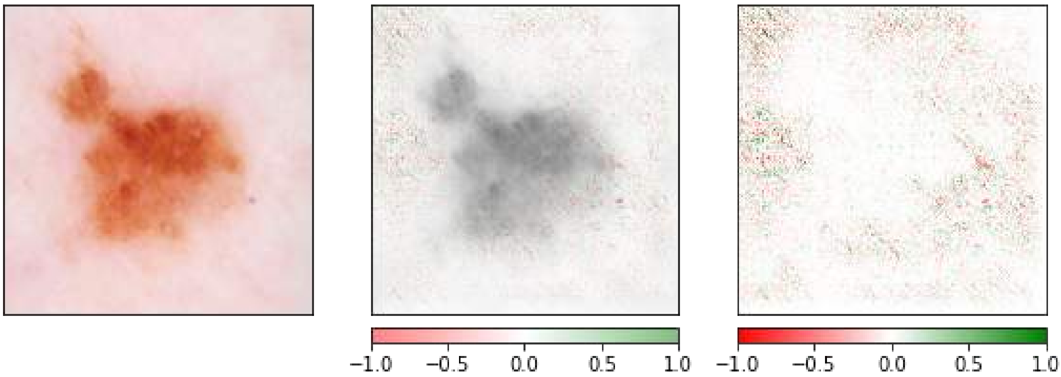
The DeepSHAP method visualizing NV dispersed dots. Left-to-right: original image, blended heat map with all values (positives and negatives), heat map with all values.

Figure 44 – DeepSHAP applied to BCC class.



The DeepSHAP method visualizing BCC lesion border and skin. Left-to-right: original image, blended heat map with all values (positives and negatives), heat map with all values.

Figure 45 – DeepSHAP applied to MEL class.



The DeepSHAP method visualizing MEL lesion border and skin. Left-to-right: original image, blended heat map with all values (positives and negatives), heat map with all values.

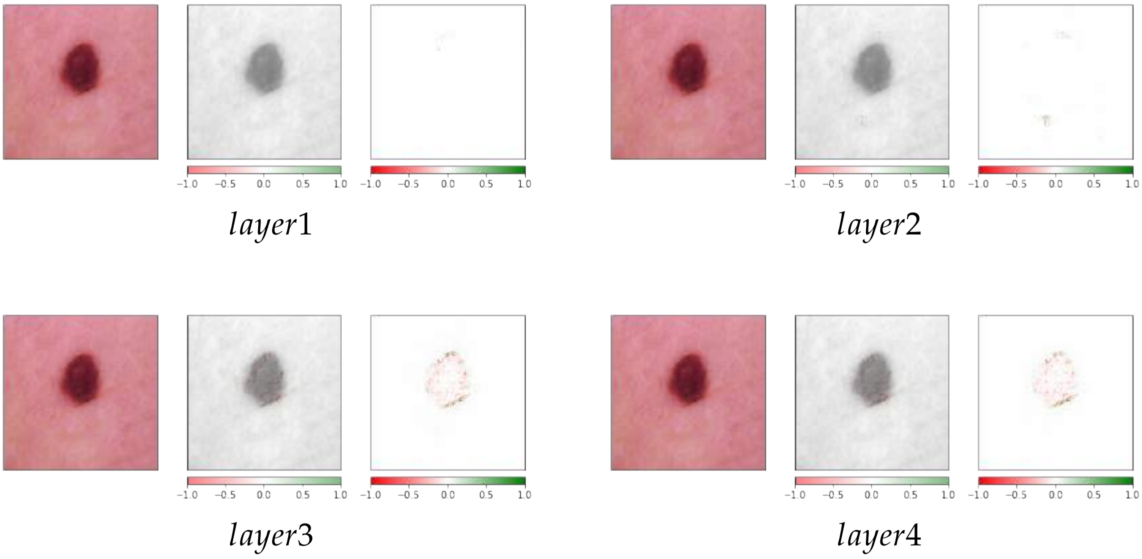
5.6 GuidedGradCam

And unlike of the other XAI methods used in this work, GuidedGradCam is used by layers of the neural network. In the case of this work, it was used in layers *layer*1, *layer*2, *layer*3 and *layer*4. As it takes some layers to better recognize the image, the results of *layer*3 and *layer*4 are easier to see where the neural network visualized to make the classification. Unlike other methods, in all cases, GuidedGradCam was able to visualize only the lesion.

The classes that had the best results in this method were AKIEC (figure 46), BKL (figure 47), DF (figure 48) and VASC (figure 49), as they visualized more points inside the lesion; in classes BCC (figure 50), MEL (figure 51), NV (figure 52), a large part of the lesion was ignored in order to predict the classification.

However, in classes BCC, BKL and NV it was also observed that the neural network focused on body hair, in BKL and NV on body hair inside the lesion and in BCC on body hair outside the lesion.

Figure 46 – GuidedGradCam applied to AKIEC class.

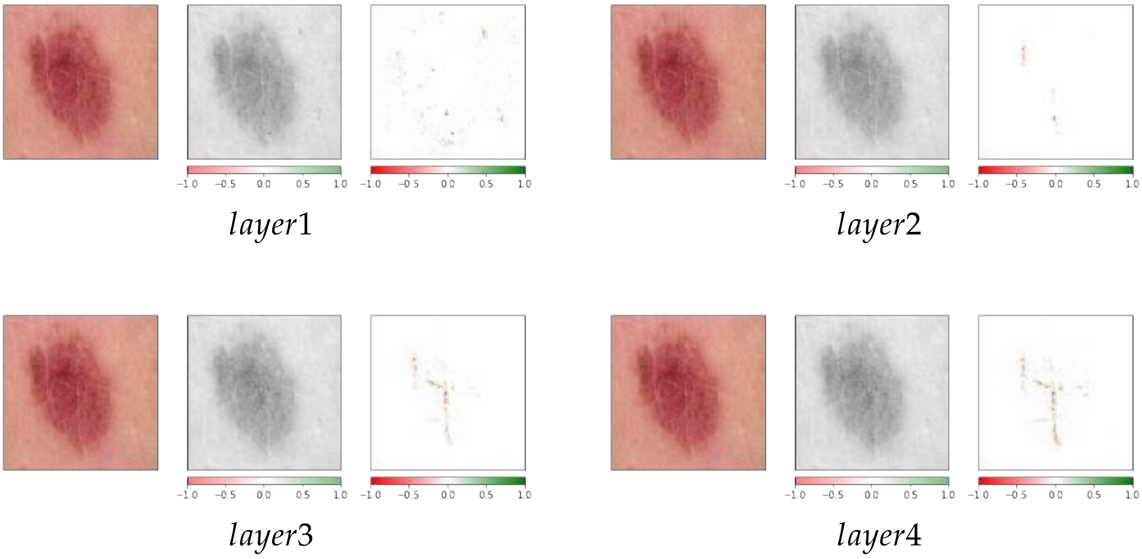


*layer*1 *layer*2

*layer*3 *layer*4

The GuidedGradCam method visualizing AKIEC lesion border with positives values and inner lesion with negative values. Left-to-right for each layer: original image, blended heat map image with all values (positives and negatives), heat map image with all values.

Figure 47 – GuidedGradCam applied to BKL class.

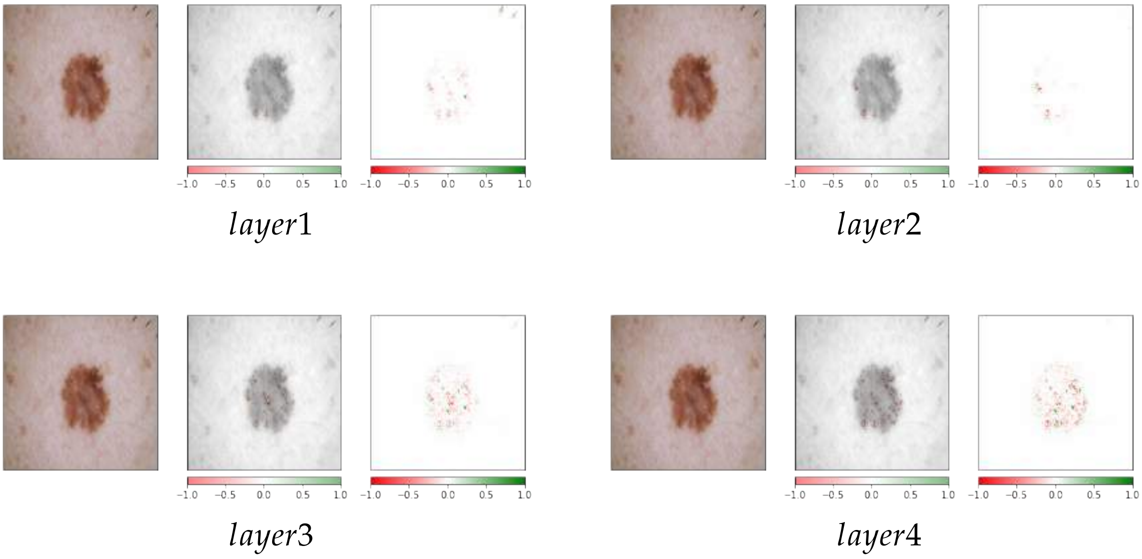


*layer*1 *layer*2

*layer*3 *layer*4

The GuidedGradCam method visualizing hair inner BKL lesion. Left-to-right for each layer: original image, blended heat map image with all values (positives and negatives), heat map image with all values.

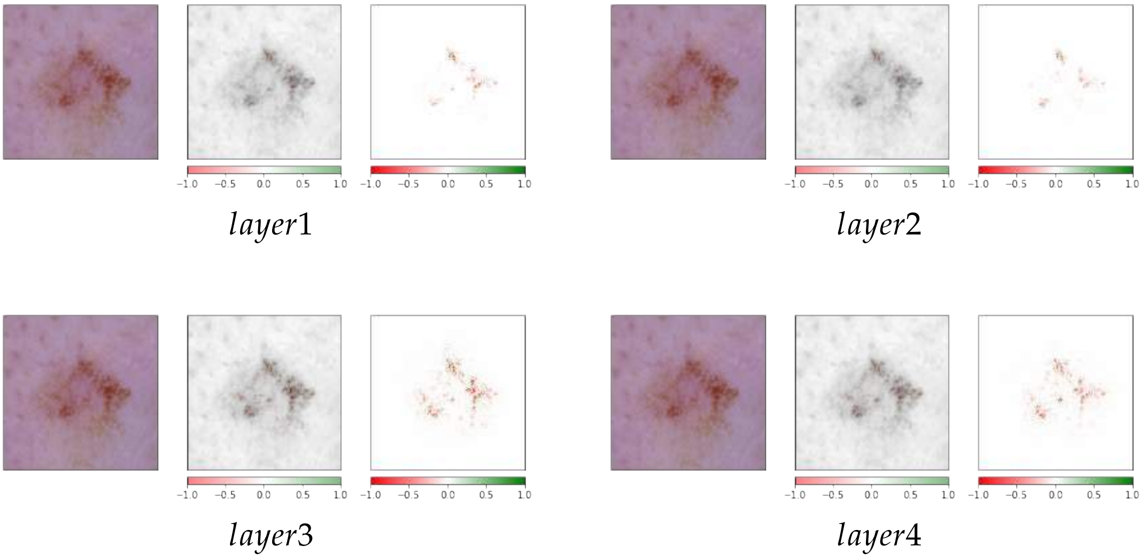
Figure 48 – GuidedGradCam applied to DF class.



*layer*1 *layer*2 *layer*3 *layer*4

The GuidedGradCam method visualizing DF lesion. Left-to-right for each layer: original image, blended heat map image with all values (positives and negatives), heat map image with all values.

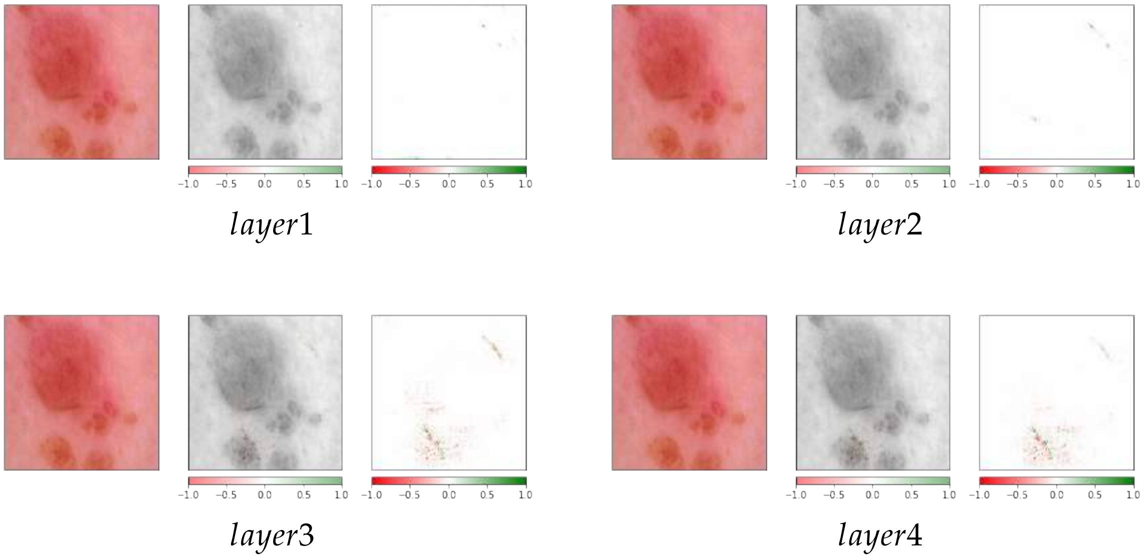
Figure 49 – GuidedGradCam applied to VASC class.



*layer*1 *layer*2 *layer*3 *layer*4

The GuidedGradCam method visualizing VASC lesion. Left-to-right for each layer: original image, blended heat map image with all values (positives and negatives), heat map image with all values.

Figure 50 – GuidedGradCam applied to BCC class.

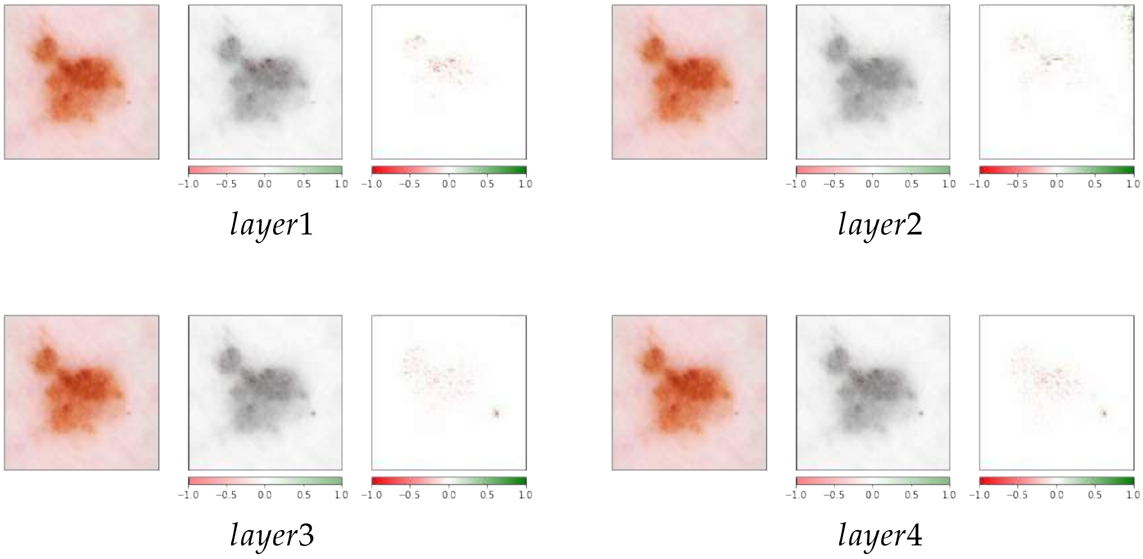


*layer*1 *layer*2

*layer*3 *layer*4

The GuidedGradCam method does not visualizing all BCC lesion and getting hair body outside lesion. Left-to-right for each layer: original image, blended heat map image with all values (positives and negatives), heat map image with all values.

Figure 51 – GuidedGradCam applied to MEL class.

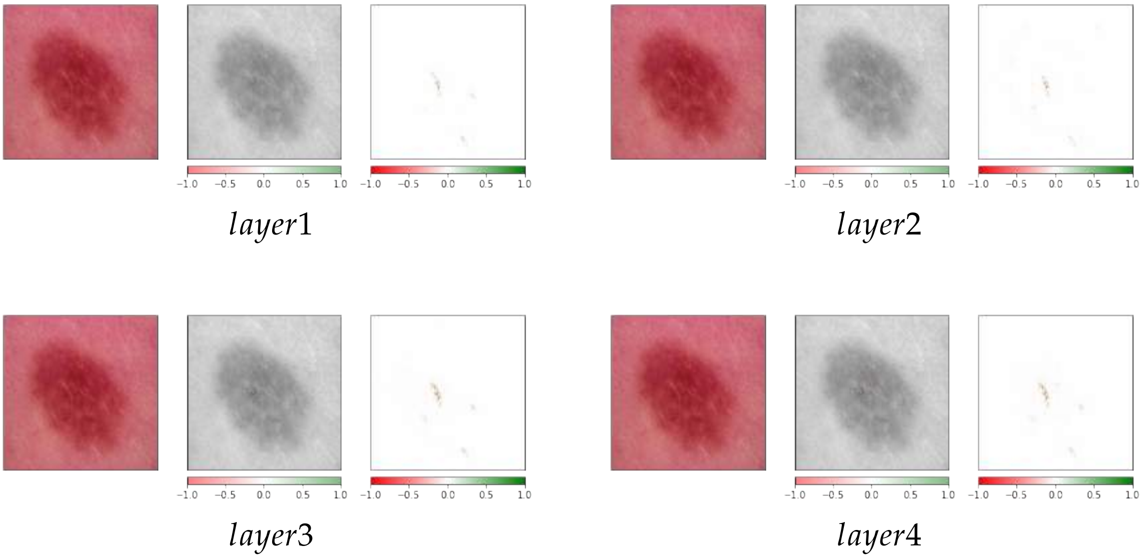


*layer*1 *layer*2

*layer*3 *layer*4

The GuidedGradCam method does not visualizing all MEL lesion. Left-to-right for each layer: original image, blended heat map image with all values (positives and negatives), heat map image with all values.

Figure 52 – GuidedGradCam applied to NV class.



*layer*1

*layer*2

*layer*3

*layer*4

The GuidedGradCam method visualizing NV lesion but few parts. Left-to-right for each layer: original image, blended heat map image with all values (positives and negatives), heat map image with all values.

5.7 Metrics Comparison

The table 7 shows the results[5] of Sensitivity and Infidelity for each XAI method used in the project for each classification used. As can also be seen, in the DeepSHAP method, no Sensitivity data was collected, due to the size of the video memory required to use the method that caused the video card's lack of video memory in the Kaggle environment.

As previously explained in the section 4.4.2, the lower the Sensitivity and Infidelity, the better the method of explanation. Interpreting the table 7, we can see that Occlusion had the best results for Sensitivity, but with the worst results for Infidelity and needing a very high time to run in each class, as shown in the table 8. Two methods that have shown promise with less sensitivity are GradientSHAP and DeepSHAP, in particular GradientSHAP for having a shorter execution time and for using less video memory.

Table 7 – Sensitivity (SENS) and Infidelity (INFD) comparison.

Lesion	DeepLIFT	Occlusion	IG	GradientSHAP	DeepSHAP	GuidedGradCAM[5]
AKIEC$_{\text{SENS}}$	0.728	**0.046**	0.705	2.802	-	0.483
BCC$_{\text{SENS}}$	0.728	**0.042**	0.662	1.807	-	0.533

Continued on next page

[5] Only layer *layer*4 for GuidedGradCAM comparison

Table 7 – Continued from previous page

Lesion	DeepLIFT	Occlusion	IG	GradientSHAP	DeepSHAP	GuidedGradCAM[5]
BKL$_{\mathbf{SENS}}$	0.613	**0.078**	0.584	1.081	-	0.291
DF$_{\mathbf{SENS}}$	0.785	**0.041**	0.683	2.602	-	0.412
MEL$_{\mathbf{SENS}}$	0.9	**0.083**	0.744	1.146	-	0.472
NV$_{\mathbf{SENS}}$	0.67	**0.084**	0.617	2.36	-	0.416
VASC$_{\mathbf{SENS}}$	0.873	**0.051**	0.705	1.036	-	0.409
AKIEC$_{\mathbf{INFD}}$	0.004	210.572	0.012	0.021	**0.005**	0.045
BCC$_{\mathbf{INFD}}$	0.007	38.939	**0.001**	0.011	0.008	0.014
BKL$_{\mathbf{INFD}}$	0.006	1.053	**0.001**	0.009	0.004	0.025
DF$_{\mathbf{INFD}}$	**0.001**	0.371	**0.001**	0.005	0.004	0.0114
MEL$_{\mathbf{INFD}}$	0.01	2.185	0.007	0.016	**0.004**	0.021
NV$_{\mathbf{INFD}}$	0.003	0.591	0.007	0.009	**0.005**	0.014
VASC$_{\mathbf{INFD}}$	0.008	5.602	**0.001**	0.011	0.008	0.016

In the table 8, which deals with the execution time, it can be observed that the GradCam method had the shortest time in comparison with the others, which was due to the GradCam execution method. As the GuidedGradCam function is used in the Captum library, it became necessary to choose which layer the method would be executed in a guided way, which in the case of this work, occurred in the 512 block layer. If you ignore the fact that GuidedGradCam stood out only in the 512 block layer ($layer4$), Integrated Gradients can be considered as the fastest method in execution, since it uses all layers of the network.

Table 8 – Execution time (s).

Lesion	DeepLift	Occlusion	IG	GradientShap	DeepShap	GuidedGradCAM[5]
AKIEC	0.846	18.24	0.816	**0.679**	0.911	0.862
BCC	0.771	17.64	0.915	**0.665**	0.896	**0.665**
BKL	0.671	18.275	0.672	**0.666**	1.057	0.669
DF	0.661	18.18	0.668	**0.659**	0.897	0.826
MEL	0.672	17.984	0.674	0.825	0.901	**0.661**
NV	**0.664**	18.245	0.82	0.666	0.901	0.669
VASC	0.673	18.404	**0.672**	0.745	1.122	**0.693**
average	0,708	18,138	0,748	**0,701**	0,955	0,721

6 Conclusion and Future Works

Taking into account the scientific need to build knowledge through advances in studies, the main objective of this work was to add knowledge through the application of the XAI methodology in convolutional neural networks in order to enable with greater reliability and interpretability the clinical analysis of images through of AI, which can represent considerable advances in the way we recognize and treat skin diseases.

Regarding the investigation of image processing, several XAI methodologies were used in order to obtain more comparison data, which in turn helped in the production of relevant results for the analysis of the work. In this sense, the use of such methodologies brought a greater understanding of where the trained ResNet-152 model was viewing the image for prediction, which is essential for the medical field, which requires greater accuracy for diagnosis. In this sense, XAI can be inferred as a promising method. Regarding the metrics, it is noteworthy that Infidelity and Sensitivity, in particular, proved to be very useful for comparing XAI methods.

As can be seen in the results section 5, the skin lesion classifications that had the best visual results were BKL and VASC, which appeared in the best results of each method XAI, NV and DF that showed good results in five of the six methods. The results of the BCC and MEL classifications, on the other hand, showed possible overffitings, as they differentiate the classifications only by the edge of the skin.

The easiest methods to see where the model was viewing were Occlusion and GuidedGradCAM, but Occlusion has a high Infidelity, especially in possible classifications that were overfitted, and has a high execution time.

The best sensitivity results were obtained using the Occlusion method, as it presented results closer to 0. The DeepLIFT, IG and GuidedGradCAM methods had good results as well. GradientSHAP achieved results above 1, reaching almost 3 in some cases, with a high sensitivity to be prone adversarial attacks. However, in the DeepSHAP method, there was difficulty in collecting Infidelity data, and for this reason, no comparison was made with this metric, as mentioned in the 5 section.

About the methods that obtained the best results on average, including Infidelity and Sensivity, were IG and DeepLIFT, respectively. At the execution time of the methods, DeepLIFT, IG and GradientSHAP had a similar average and had the best results. Occlusion, on the other hand, obtained a very high execution time compared to the other methods.

For the purposes of continuing the Mendes and Silva (2018) research, the AUC data had greater results than the previous work with the use of the HAM10000 dataset, with the exception of melanoma and naevus, and it should be noted that the latter presented results for below when compared to studies of Mendes and Silva (2018) and Han et al. (2018), this factor may have happened due to the use of another dataset.

For future work, the use of these XAI methods in other architectures (such as Hourglass Network) is suggested, applying the HAM10000 dataset and comparing them with more XAI post-hoc (such as LIME and Saliency) or ante-hoc (such as Bayesian Deep Learning) methods, as this may represent, as in the present research, advances in the applicability and usability of the methodology in this field, allowing an optimization of knowledge.

Finally, considering the importance of greater integration of scientific data and broad consent in the literature on the relevance of the XAI theme, the use of datasets with a greater repertoire of samples in different skin tones is also suggested, in order to match the studies in XAI methodologies in image processing of skin lesions to the reality of human diversity in this aspect. This need materializes in the present study, as this factor was not further explored due to the use of the HAM10000 dataset, which had more images of lesions recorded on Caucasian skin. In this sense, the greater diversity of data could improve the accuracy of classifications, especially in classifications that focused more on the skin than on the lesion itself.

7 Bibliography

ANCONA, M. et al. A unified view of gradient-based attribution methods for deep neural networks. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, abs/1711.06104, 2017. Available at: <http://arxiv.org/abs/1711.06104>.

ARRIETA, A. B. et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. oct 2019. Available at: <http://arxiv.org/abs/1910.10045>.

BISSOTO, A. et al. (De)Constructing Bias on Skin Lesion Datasets. *IEEE Xplore*, n. Ic, 2019. Available at: <http://arxiv.org/abs/1904.08818>.

BLOICE, M. D.; STOCKER, C.; HOLZINGER, A. Augmentor: An image augmentation library for machine learning. *CoRR*, abs/1708.04680, 2017. ISSN 23318422. Available at: <http://arxiv.org/abs/1708.04680>.

CHOI, E. et al. RETAIN: interpretable predictive model in healthcare using reverse time attention mechanism. *CoRR*, abs/1608.05745, 2016. Available at: <http://arxiv.org/abs/1608.05745>.

ESTEVA, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, Nature Publishing Group, v. 542, n. 7639, p. 115–118, 2017. ISSN 14764687. Available at: <http://dx.doi.org/10.1038/nature21056>.

FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters*, v. 27, n. 8, p. 861–874, 2006. ISSN 01678655. Available at: <https://doi.org/10.1016/j.patrec.2005.10.010>.

FERRUCCI, D. et al. Watson: Beyond jeopardy! *Artificial Intelligence*, v. 199-200, p. 93–105, 2013. ISSN 0004-3702. Available at: <https://www.sciencedirect.com/science/article/pii/S0004370212000872>.

GIOTIS, I. et al. MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert Systems with Applications*, Elsevier Ltd, v. 42, n. 19, p. 6578–6585, 2015. ISSN 09574174. Available at: <http://dx.doi.org/10.1016/j.eswa.2015.04.034>.

HAJIAN-TILAKI, K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med*, v. 4, n. 2, p. 627–635, 2013.

HAN, S. S. et al. Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. *Journal of Investigative Dermatology*, The Authors, v. 138, n. 7, p. 1529–1538, 2018. ISSN 15231747. Available at: <https://doi.org/10.1016/j.jid.2018.01.028>.

HE, K. et al. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. Available at: <http://arxiv.org/abs/1512.03385>.

HOLZINGER, A. et al. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, v. 9, n. 4, p. e1312, 2019. Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1312>.

KAUL, V.; ENSLIN, S.; GROSS, S. A. History of artificial intelligence in medicine. *Gastrointestinal Endoscopy*, v. 92, n. 4, p. 807–812, 2020. ISSN 0016-5107. Available at: <https://www.sciencedirect.com/science/article/pii/S0016510720344667>.

KIM, T. W. Explainable artificial intelligence (XAI), the goodness criteria and the grasp-ability test. p. 1–7, oct 2018. Available at: <http://arxiv.org/abs/1810.09598>.

KOKHLIKYAN, N. et al. Captum: A unified and generic model interpretability library for pytorch. *CoRR*, abs/2009.07896, 2020. Available at: <https://arxiv.org/abs/2009.07896>.

KULIKOWSKI, C. A. An Opening Chapter of the First Generation of Artificial Intelligence in Medicine: The First Rutgers AIM Workshop, June 1975. *Yearb Med Inform*, v. 10, n. 1, p. 227–233, Aug 2015.

KUMAR, R.; INDRAYAN, A. Receiver operating characteristic (ROC) curve for medical researchers. *Indian Pediatrics*, v. 48, p. 277—287, 2011.

LIN, Z. Q. et al. Do Explanations Reflect Decisions? A Machine-centric Strategy to Quantify the Performance of Explainability Algorithms. p. 1–9, 2019. Available at: <http://arxiv.org/abs/1910.07387>.

LUCAS, R.; TONY, M.; ARMSTRONG, S. W. B. *Solar ultraviolet radiation : global burden of disease from solar ultraviolet radiation*. [S.l.]: World Health Organization, Public Health and the Environment, 2006. 250 p. ISBN 9241594403.

LUNDBERG, S.; LEE, S. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, n. Section 2, p. 4766–4775, 2017. ISSN 10495258. Available at: <http://arxiv.org/abs/1705.07874>.

MENDES, D. B.; SILVA, N. C. da. Skin Lesions Classification Using Convolutional Neural Networks in Clinical Images. dec 2018. Available at: <http://arxiv.org/abs/1812.02316>.

NATIONS, U. *MULTILATERAL Montreal Protocol on Substances that Deplete the Ozone Layer (with annex)*. 1989. Available at: <https://treaties.un.org/doc/publication/unts/volume%201522/volume-1522-i-26369-english.pdf>.

ORGANIZATION, W. H. *WHO INFORMATION SERIES ON SCHOOL HEALTH DOCUMENT SEVEN Sun Protection: An Essential Element of Health-Promoting Schools*. 2002. Available at: <https://apps.who.int/iris/bitstream/handle/10665/67400/WHO_NPH_02.6.pdf>.

PARKER, E. R. The influence of climate change on skin cancer incidence – a review of the evidence. *International Journal of Women's Dermatology*, Elsevier, Jul 2020. Available at: <https://www.sciencedirect.com/science/article/pii/S2352647520301155>.

PASZKE, A. et al. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703, 2019. Available at: <http://arxiv.org/abs/1912.01703>.

PEREZ, L.; WANG, J. The effectiveness of data augmentation in image classification using deep learning. *CoRR*, abs/1712.04621, 2017. Available at: <http://arxiv.org/abs/1712.04621>.

RAI, S. Benign skin lesions. *Medicine (United Kingdom)*, Elsevier Ltd, v. 45, n. 7, p. 435–437, 2017. ISSN 13654357. Available at: <http://dx.doi.org/10.1016/j.mpmed.2017.04.008>.

SCHLEGEL, U. et al. Towards a Rigorous Evaluation of XAI Methods on Time Series. n. Ml, 2019. Available at: <http://arxiv.org/abs/1909.07082>.

SELVARAJU, R. R. et al. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. Available at: <http://arxiv.org/abs/1610.02391>.

SHRIKUMAR, A.; GREENSIDE, P.; KUNDAJE, A. Learning important features through propagating activation differences. *34th International Conference on Machine Learning, ICML 2017*, v. 7, p. 4844–4866, 2017.

Sociedade Brasileira de Dermatologia. *O que é cancer de pele?* 2017. Available at: <https://www.sbd.org.br/dermatologia/pele/doencas-e-problemas/cancer-da-pele/64/>.

SUNDARARAJAN, M.; TALY, A.; YAN, Q. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365, 2017. Available at: <http://arxiv.org/abs/1703.01365>.

SWETS, J. A. Indices of Discrimination or Diagnostic Accuracy. Their ROCs and Implied Models. *Psychological Bulletin*, v. 99, n. 1, p. 100–117, 1986. ISSN 00332909.

TSCHANDL, P. *The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions*. Harvard Dataverse, 2018. Available at: <https://doi.org/10.7910/DVN/DBW86T>.

World Health Organization. *Helping people reduce their risks of skin cancer and cataract*. World Health Organization, 2002. Available at: <https://www.who.int/news/item/22-07-2002-helping-people-reduce-their-risks-of-skin-cancer-and-cataract>.

World Health Organization. *The Global UV Project A Guide and Compendium*. 2003.

World Health Organization. *The World Health Organization recommends that no person under 18 should use a sunbed*. World Health Organization, 2005. Available at: <https://www.who.int/news/item/17-03-2005-the-world-health-organization-recommends-that-no-person-under-18-should-use-a-su

World Health Organization. *Artificial tanning devices: Public health interventions to manage sunbeds*. World Health Organization, 2017. Available at: <https://www.who.int/publications/i/item/9789241512596>.

YEH, C. et al. How sensitive are sensitivity-based explanations? *CoRR*, abs/1901.09392, 2019. Available at: <http://arxiv.org/abs/1901.09392>.

ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013. Available at: <http://arxiv.org/abs/1311.2901>.