Interpretable Diagnosis of Skin Cancer Using Deep Learning

Samara S. Santos [*,**,***] **Tamires M. Rezende** [***]
Marcos A. Alves [**,***] **Frederico G. Guimarães** [**,****]

[*] *Federal Center for Technological Education of Minas Gerais (CEFET-MG), Timóteo, Minas Gerais, Brazil*
[**] *Graduate Program in Electrical Engineering, Federal University of Minas Gerais, Av. Antônio Carlos 6627, Belo Horizonte, Minas Gerais, Brazil (e-mail: {samarass,marcosalves,fredericoguimaraes}@ufmg.br).*
[***] *Machine Intelligence and Data Science (Minds) Laboratory, Federal University of Minas Gerais, Av. Antônio Carlos 6627, School of Engineering, Block I, Room 2200, Belo Horizonte, Minas Gerais, Brazil. (e-mail: inc.tamires@gmail.com)*
[****] *Department of Electrical Engineering, Federal University of Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil*

Abstract: Skin cancer is the main type of cancer that affects people all over the world, being melanoma the most feared, due to its rapid spread throughout the body. If it is detected in the early stages, the chances of cures are above 96%. Due to this, approaches to help the clinicians in the correct diagnosis, as well as that focused on the explanations, have been largely explored. In this context, this paper aims to build a binary classification of skin moles model using the ResNet50 and explain its prediction by comparing two known explainer tools LIME and Decision Trees (DT). The ResNet50 architecture achieved results of about 92% in terms of accuracy and up to 91% in sensibility and specificity. When LIME and DT were compared, both showed no fidelity error. However, in terms of stability, measured by the Jaccard index, LIME presented an stability of 0.497 ± 0.473 and DT of 1.0 ± 0.0, showing stability only for the latter. These results were obtained from 30 runs of images randomly chosen from the test base. Through a visual analysis, LIME varied in two of the 5 images from the benign and in one from the malignant lesion. As important as generating good classification models is providing clinicians with good explanation models that are intuitive and consistent.

Keywords: Skin Cancer Diagnosis; ResNet50; XAI; LIME; Decision Tree; Interpretability.

1. INTRODUCTION

Although medical science has been going forward in the last centuries most than ever, disease prognostics is still a current demand in healthcare. For this, many Machine Learning (ML) approaches for medical diagnostic have been applied using pattern recognition techniques (Tschandl et al., 2019). One of them is the skin cancer diagnostics from a set of images of the skin lesion, which is the focus of this work.

The skin is the longest organ in the human body which protects the body against heat, light, and infection, besides helping to control the body temperature and to store the fat and the water (Zhang et al., 2020). Skin cancer is the most common type of cancer, responsible for 70% of the diagnostics (Tschandl et al., 2019). In this sense, skin cancer preventive detection is a key role to anticipate people's treatment, and also prevent some variants, such as focal cell carcinoma and melanoma (Zhang et al., 2020).

This pathology is primarily diagnosed through visual inspection, starting from initial clinical screening to dermoscopic analysis, biopsy, and histopathological examination (Esteva et al., 2017). The most aggressive skin cancer is melanoma since it can spread quickly to other organs. Occasionally, it also spreads through the lymphatic or circulatory system and can achieve the farthest points of the body. However, the percentage of the cure for those who have early diagnosis varies between 96% to 99% (Esteva et al., 2017).

Jaleel et al. (2013) mentioned some inconveniences in dealing with melanoma: (i) the first wound of the disease can be the path for new ones; (ii) the biopsy method may cause inflammation or even spread of lesions; and (iii) there is a similarity between benign and malignant melanomas which takes more attention by the clinicians. Although the diagnosis made by the doctor is reliable and valuable, there is a need for non-invasive approaches. However, many of them take lots of time and require high cognitive efforts, while the computer vision-based approaches can be used to help the medical team in having a more accurate clinical diagnosis (Mehta and Shah, 2016).

ISSN: 2525-8311

DOI: 10.20906/CBA2022/3215

The commonest rule used by specialists to classify the lesion as benign or malignant is named ABCDE, or Asymmetry, Border, Color, Diameter, and Evolving. Asymmetry refers to the shape of the lesion when half of the lesion does not match the other one. Also, irregular borders and colors indicate the possibility of melanoma diagnosis. Regards the size of the lesion, generally, melanoma is greater than 6mm. Then, anomalies in one or more of those parameters indicate the malignancy of the lesion.

Similarly, the application of the Convolutional Neural Networks (CNN) has presented prominent results in solving this kind of problem-based on the ABCDE and/or other rules. Sometimes the result obtained surpasses the results of specialist doctors (Brinker et al., 2019; Esteva et al., 2017). However, misclassifications in skin cancer diagnostics may lead to serious clinical consequences, delaying the treatment or even more complicated issues. Due to this, there is a need for a better understanding of the results from the CNN classifiers, which is directly related to Explainable Artificial Intelligence, or simply XAI.

The XAI help to understand the regions of the images that are significant for the predictions, and consequently leads to a confident result by the specialist point of view (Van Der Velden et al., 2022; Alves et al., 2021; Ferreira et al., 2020; Santos et al., 2021). The Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016; Yu, 2018) and Decision Trees (DT) (Loh, 2011) have been used for this purpose. They allow the specialist a better understanding of the reasons that justify the prediction made by the learning method.

It is known that most existing XAI focus on structured data. However, some methodologies can help in the interpretation of deep networks. According to Van Der Velden et al. (2022), saliency mapping is the most common form of XAI in medical image analysis since it shows the important parts of an image for a decision, often based on back-propagation techniques. Perturbation-based approaches can be also mentioned. It creates new instances by perturbing input data and then, measuring how perturbed input data changes the output, as in LIME. Both of them highlight the image's area important to the prediction. By knowing that, the specialists may evaluate if the model may capture the relevant knowing information.

With this in mind, this work proposes a quantitative and visual comparison of the explanations provided by LIME and DT for the classification of skin cancer lesions. Although it is widely used for image interpretation, some works show that when applied to structured data, LIME may have low local fidelity and stability for its explanations (Ferreira et al., 2020; Santos et al., 2021). Thus, we propose to investigate whether these problems are also observed in the interpretation of deep learning and whether when the internal model used to adjust the data is changed, from a simple linear model to a DT, if significant changes are observed in the evaluated metrics.

Additionally, another aspect worth mentioning is that in the last few years, many regulations for data usage and IA systems were proposed around the world. The Europe Union have proposed the General Data Protection Regulation (GDPR) to regulate and protect their people from the effects of data processing and AI decisions. In Brazil, since 2020 the *"Lei Geral de Proteção de Dados"* (General Data Protection Law, in free translate) with similar objectives to GDPR (Santos et al., 2021). In this way, the XAI aims to reinforce the ways to achieve explainability and interpretability for image data.

The rest of the paper is organized as follows: Section 2 describes related works pointing out those focused on the automatic diagnosis of skin cancer. Section 3 details the materials, methods and experiments. Section 4 describes and discuss the results. Finally, Section 5 concludes the paper and presents directions.

2. RELATED WORKS

Scientific advances have made it possible to improve the equipment that helps in the diagnosis of skin cancer, but they still may fail. Computerized techniques, such as CNN, serve as an important support tool for medical diagnosis since many of these systems were able to surpass human specialists in this diagnosis (Tschandl et al., 2019). The joint application of these two experts (dermatologist + specialist systems) aids to bring more confidence to the diagnosis, reducing the cost of diagnosis and the complications that may occur when the patient is submitted to specific exams.

Different systems were proposed to deal with that problem. Some of them are based on the traditional ML models (Alves et al., 2021; Vidya and Karki, 2020; Javaid et al., 2021), while other ones apply DL or ensemble methods (Thurnhofer-Hemsi and Dominguez, 2020; Daghrir et al., 2020).

Javaid et al. (2021) and Vidya and Karki (2020) proposed a binary classification for a skin lesion image input using data from ISIC-ISBI 2016 repository. Although they used different ML methods, the former achieved 93.89% of accuracy with a model built with the Random Forest (RF) algorithm and the latter 97.8% with Support Vector Machines (SVM).

Different from the traditional methods, DL can also be used to solve the skin cancer diagnosis problem without using an image preprocessing task. Thurnhofer-Hemsi and Dominguez (2020) implemented a multi-class problem with seven classes: Actinic Keratoses, Basal cell carcinoma, Benign keratosis, Dermatofibroma, Melanoma, Nevi, and Vascular skin using the HAM10000 dataset. The main concern with this dataset is that almost 70% of the images are of the Nevi class, which means that the data is highly unbalanced. To circumvent this issue, the authors proposed a two-level architecture, named DenseNet2021. The first layer separates the Nevi class from the others, while the second one classifies the other six classes. It was reported a performance 10% better than the literature.

Another approach was proposed by Daghrir et al. (2020) using a hybrid approach, combining the result of CNN, KNN, and SVM methods. The authors presented a new concept named "ugly duckling" regarding outliers data. In the preprocessing stage, the hairs were removed from the images, then the OTSU segmentation, and the ABCD, and Blue-Black rules are applied. The final accuracy of 88.4% was obtained from the majority vote of the three methods, which overcomes each method's results. Hosny et al. (2018)

developed an architecture to classify color images of skin cancer into three classes: Melanoma, atypical nevus, and common nevus. As stated by the authors, this method has the advantage over earlier computer-aided methods for this purpose because it can work with any type of image (dermoscopic and photographic), besides it is not necessary for any preprocessing, since it works directly with the colored skin images.

Regarding interpretable systems, Thomas et al. (2021) proposed interpretable deep learning methods for multi-class segmentation and classification of non-melanoma skin cancer, in which outputs of the network are present in an interpretable way and can be visualized in several forms to distinguish its capabilities. Jiang et al. (2021) introduced a new deep learning-based approach for the automated diagnosis of skin cancer, named “DRANet” and a “Class Activation Map” (CAM) to obtain visual explanations from the deep neural network. As argued by the authors, “the CAM can convert the output of attention modules to a heat map showing key areas where the model focuses more”. Xie et al. (2019), in its turn, the visualization of CNN representations is present to identify cells between melanoma and nevi.

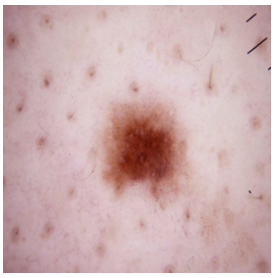
As can be seen, although many classic ML and DL techniques have been used to solve this problem, most of them do not mention the use of transparent solutions or techniques to interpret the results achieved. As a result, *post-hoc* methodologies, which can be applied after the system is ready, greatly contribute to the system’s reliability as well as to its implementation and maintainability. Examples of LIME applications for generating explanations of individual predictions of DL models in the medical field include Parkinson’s disease diagnosis Magesh et al. (2020), ophthalmology (Hanif et al., 2021), and the proper skin lesion diagnosis (Xiang and Wang, 2019). DTs, outside the DL context, have been used for a local explanation as in Alves et al. (2021). These authors proposed a local DT Explainer, or “DTX”, for COVID-19 detection.

This paper applies the known Residual Neural Network (ResNet50) (He et al., 2016) in the ISIC-Archive dataset to make the binary classification of skin cancer (benign or malignant). For providing the explanations provided by the proposed ResNet50 architecture, it was used LIME and DT explainers and evaluated in terms of fidelity and stability.

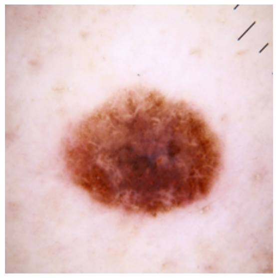
3. MATERIAL AND METHODS

3.1 Dataset

The ISIC-Archive dataset, publicly available at Kaggle (2019), contains a balanced dataset of images of benign and malignant skin moles. The data is present in two folders with 1800 pictures (224×244) of the two types of lesions, labeled and previously divided into training and testing data. However, for evaluation purposes, it is going to be redivided using the re-sampling technique named k-fold cross-validation. Figure 1 presents an example of benign and malignant lesions from this dataset. Since the images are already the proper size for deep net training, resizing them is not necessary.



(a) Benign lesion



(b) Malignant lesion

Figure 1. Examples of a skin lesion in the ISIC-Archive dataset
Source: Kaggle (2019)

3.2 RestNet50 Architecture

The ResNet50 (He et al., 2016) is an artificial neural network based on pyramidal cell constructs in the cerebral cortex. This network was proposed in a context where the deep networks known so far had a limitation in increasing the number of layers. It was observed that after a certain number of internal convolutional layers, the training and test error start to increase since the error back-propagation does not reach the first layers.

To build this network, convolutional layers are grouped employing the so-called ResBlock, illustrated in Figure 2, through “skip connections” that help in back-propagating the error, and consequently avoiding the gradient explosion. These connections provide an alternative gradient shortcut path to flow through. Also, they allow the model to learn an identity function that guarantees that the top layer will outperform the bottom layer at least as well, and no worse. This model was the winning entry in the ImageNet competition in 2015. The key advance with ResNet was enabling the training of extremely deep neural networks with more than 150 layers.

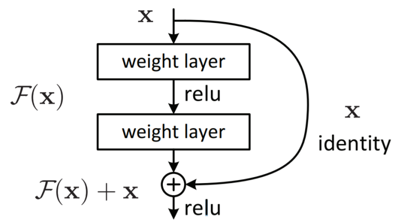


Figure 2. ResBlock
Source: He et al. (2016)

The ResNet has variants with different numbers of convolutional layers. The used notation used is the name ResNet followed by a number that specifies how many layers the network has. Here, we choose the ResNet-50 variant, in virtue of the good performance of this architecture to solve image recognition problems (Fulton et al., 2019).

The network takes the input image with height, width, and 3 channels. Initial convolution is performed applying 7×7 and 3×3 kernel sizes, respectively, followed by the maximum clustering. The ResNet50 architecture has 4 stages: stage 1 has 3 residual blocks with 3 layers each, and the kernel sizes used to perform the convolution operation on all 3 layers of the block in this stage are 64, 64, and 128

ISSN: 2525-8311

DOI: 10.20906/CBA2022/3215

respectively. From one stage to the next, the input size is halved while the channel width is doubled.

In addition, for each residual function, 3 layers are stacked on top of each other. The three layers are 1×1, 3×3, 1×1 convolutions. The 1×1 convolution layers are responsible for reducing and restoring the dimensions. The 3×3 layer is left as a bottleneck with smaller input/output dimensions (Sachan, 2019). In the end, average pooling is applied followed by a fully connected layer with 1000 neurons.

3.3 Network Interpretability with LIME and DT

The Local Interpretable Model-agnostic Explanations (LIME) is a framework that explains how the input attributes of an ML model impact its output predictions. In the case of image classification, LIME determines the sub-regions of an image (set of superpixels) with the strongest association with a prediction label. To generate explanations for a black-box model for a specific instance, LIME creates a new dataset from random perturbations (with their respective outputs) around the sample to be explained and then fits a weighted local surrogate model. This local model is usually a simpler model with intrinsic interpretability such as a linear regression model. It generates explanations with the following steps:

(1) Generate random perturbations around the image to be explained;
(2) Obtain the predictions of the DL model for perturbations;
(3) Compute the importance of the perturbations (for each super-pixel, which represented the features of the image);
(4) Fit is an explainable linear model using the perturbations, predictions, and weights.

Using these steps it is possible to verify which parts of the lesion were the most important for the prediction as benign or malign.

Decision Trees (DT), in its turn, are widely used statistical methods for classification and regression tasks. This method partitions the feature space recursively into subregions, from an impurity function. In a DT, each of the internal nodes, including the root node, represents a hyperplane that is parallel to the attribute axis used for this partitioning. The leaf nodes, on the other hand, represent the classes of the problem, so that the path between the root node and the leaf node used to make a prediction is easily identified.

Due to this, this approach is remarkably interpretable, and by using the same strategies of LIME, evaluating the vicinity of the sample of interest, may be applied to generate local explanations. The main difference between these two approaches concerns the adjustment of the interpretable method since LIME uses a simple linear model and DT uses the structure itself to provide explanations.

For sake of simplicity, an overview of the proposed methodology is shown in Figure 3.

3.4 Evaluation Metrics

The explanation was evaluated in terms of fidelity and stability. Fidelity refers to how well the surrogate method can replicate the behavior of the black-box method locally. Given the perturbation of the superpixels, which generates new samples, and the respective prediction value, the percentage of correct answers of LIME and DT for each perturbation i is verified and compared with the value generated by the black box. In other words, the learning error of the surrogate method is verified by checking the behavior of the opaque ML. Mathematically, this concept may be understood equals to the known Mean Squared Error (MSE) between the probability prediction made for the model and the explainer, as presented in Equation (1).

$$err(f) = \frac{1}{N}\sum_{i=1}^{N}(y_i^{'} - y_i)^2 \tag{1}$$

where f represents the approximation function between the model and the explainer. y' is given by the surrogate model and real value (y) for sample i, N the number of samples evaluated.

Stability is expected for different runs of the explainer applied to the same sample and in the same conditions to generate the same explanation. To measure it, the Jaccard index was used. It is a statistical measure used to assess the similarity and diversity of sample sets. Thus, this index evaluates the stability of the resource selection process, since it is expected that the selected resources are similar, meaning that the selection process is consistent. The first metric is calculated from the intersection performed on sets of two or more runs, divided by the union between these two sets, according to Equation (2).

$$J(A,B) = \frac{A \cap B}{A \cup B} \tag{2}$$

where A and B are two lists of the most important features obtained in different executions using the same parameters.

To know, the best possible value for approximation error (or fidelity) is 0 meaning that there is no error between the CNN predictions and the surrogate model. And the best value for the Jaccard index is 1, which means that the model maintains stability when returning the most important features.

3.5 Experimental Procedures

The code for the experimental procedure was developed in Python, using the Keras package from TensorFlow (Gulli and Pal, 2017). The network input images were already sized appropriately for the input of RestNet50, from the 224×224. The normalization was carried out in the proportion of 1/255. Data augmentation was performed at run-time by applying the “ImageDataGenerator” function that generates more specimens of both classes and consequently identifies the relevant characteristics to differentiate the characteristics of one lesion from another. Data augmentation aims to build new instances through data transformations. It was tested at different rotations, but in the end, 20 degrees was used.

The batch size was defined as 34, and the training was performed with 30 epochs. The convolutional layers were configured with 64 filters and stride of 3 × 3 size. The dropout rate was defined as 0.3. The pooling applied was Max polling, with size 2 × 2, and the Dense layers

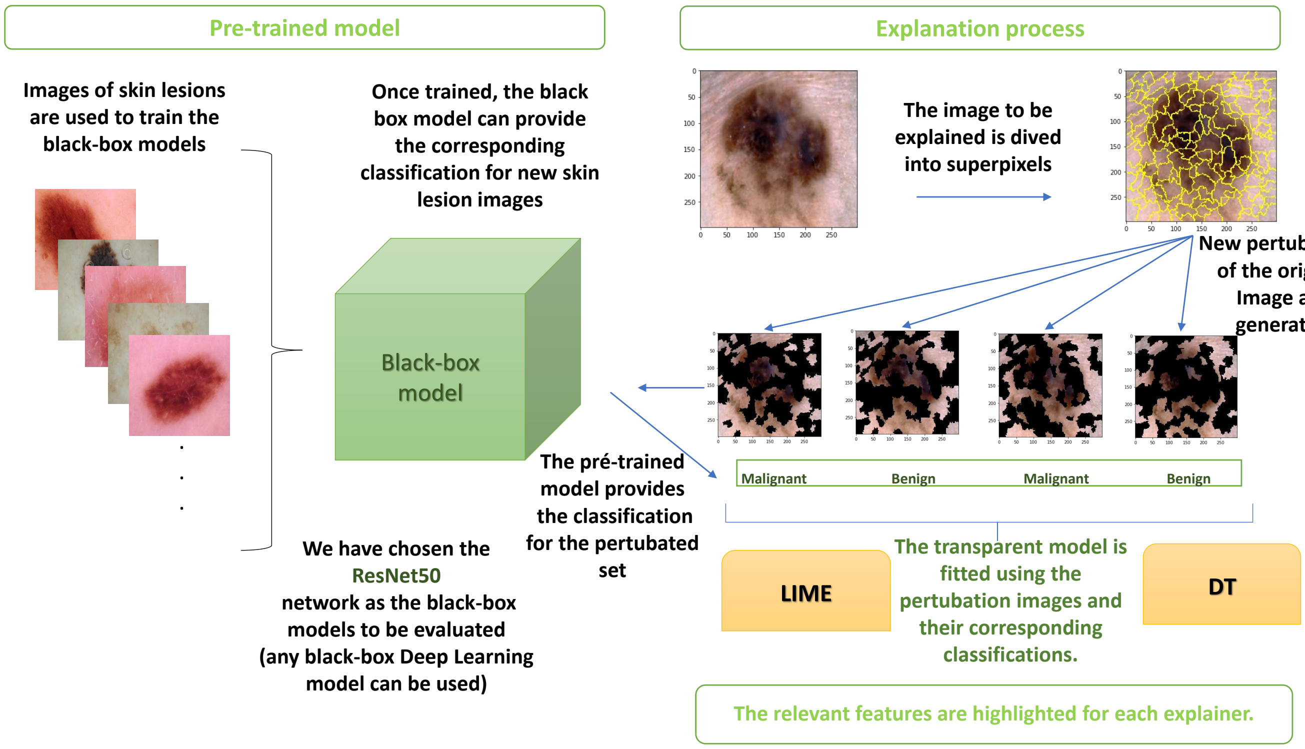


Figure 3. Overview of the methodology adopted considering the (i) generation of a vicinity around a noise sample to be explained, (ii) to obtain the predictions from the ML black-box model for the perturbed images, and (iii) the fitted of these images with LIME and DT. With the transparent models, it is possible to explain the test case.

were activated by the Relu function, except the last one, where was applied softmax function was. The filter size was applied to follow the original definitions proposed by He et al. (2016). The learning rate chosen was $lr = 0.00001$, with Adam optimization and cross-entropy as loss function.

In short, the ResNet50 model from Keras library was used, with image width and height of 224 x 224, a sequential model, with a convolutional layer with 64 input filters and a 3x3 dimension. Next, a dropout layer was added with a rate equal to 0.3. Another convolutional layer with a parametric configuration similar to the previous one. A 2×2 sized pooling max tier, plus a dropout tier at the same rate of 0.3. Two dense layers with 512 and 256 filters, respectively, plus another dropout layer. Six more dense layers, with filter multiples of 2 ranging from 128, 64, ..., 2, with all layers except the last one using the Relu activation function, and the last layer doing the binary classification applying a softmax.

The settings shown were obtained after variations in image size and various data augmentation operations. The parameters mentioned above correspond to the best configuration for the ABCD rule.

For LIME and DT, the superpixels were generated using the Quickshift segmentation method, with kernel size equal to 3 and ratio 0.4, which were parameterized using the *skimage* package (Van der Walt et al., 2014). A set of 200 new samples were generated around the image of interest (perturbations of $\mathbf{x}$), based on the binomial distribution, with $n = 1$, $p = 0.5$, being n the number of trials, and p the probability of success.

The samples to be explained were randomly picked from the test set, with a 50% probability of generating a sample that corresponds to a benign lesion, and 50% for malignant.

Fidelity and stability refer to 60 executions, both for LIME and DT. To compute stability, for each random sample, the explanation method is run 2x using the same hyperparameters, generating two lists of best attributes, following the Jaccard index. Fidelity is computed only once for each sample. Therefore, a total of 60 runs were performed, resulting in 30 values for stability and fidelity for each method of explanation.

4. RESULTS

The increase in data with the shift in the width and height of the image was removed since one of the aspects assessed by specialists when applying the ABCD rule considers the diameter of the lesion, and data augmentation with noise and contrast application has been added. However, no improvements were observed in relation to the accuracy values obtained in the training and test bases. Thus, other parameters were tested, and the change in the learning rate showed differences in the final result when only rotations are performed to increase data (without applying noise, contrast, and shifts). In addition, 20 epochs in total were performed, and rotation was also changed, from 45º to 20º. The results are presented in Table 1.

ISSN: 2525-8311

DOI: 10.20906/CBA2022/3215

Table 1. ResNET50 performance

Training	Test	Sensibility	Specificity
0.999 ± 0.002	0.921 ± 0.006	0.914 ± 0.015	0.917 ± 0.001

The Table 2 shows the results obtained for the stability and fidelity of the evaluated explainers. It was applied Mann-Whitney U Test. It was found that the stability has significant mean differences in the level of 95% of confidence.

Table 2. Interpretability methods evaluation

Method	Stability
LIME	0.497 ± 0.473
DT	1.0 ± 0.0
z-score	3.54087
p-value	0.0004

Two images from the test set were chosen randomly for evaluating the stability of the LIME and DT: one representing a benign lesion and the other a malignant one. Figure 4 illustrates the explanations provided by LIME and DT for an image that represents a type of benign skin cancer lesion (left). Through visual analysis of the benign lesion it is possible to note that the LIME was not consistent for two of the five samples presented (inside the red line), while the Decision Tree showed the same explanations in each run of the experiment.

Figure 5 also illustrates the explanations for LIME and DT, but now for a malignant lesion, shown on the left. In this case, LIME showed different results in the explanation for one of the five images, indicated inside the red line. DT, as in the previous case, was consistent for the five examples.

As argued in the literature, explanation models should be used as decision support tools by the experts helping them in their daily activities. DTs, in this case, showed more promising results than LIME through the visual analysis. It is known that decision trees can present the decisions themselves for the final prediction. In practical terms, it can be very interesting to evaluate the final DT decision based on internal decisions. This could also support the classification outcome and assist readers in understanding the explainable process. This is a very interesting reason for applying decision trees for explainability purposes. Finally, it is important to evaluate the correlation between visual insights provided by experts and visual insights provided by XAI techniques to analyze the match between the accuracy level and agreement between computational and manual segmentation.

5. CONCLUSIONS

Skin cancer diagnosis through images plays an important role in the treatment of the disease. Since the initial analysis is often done with a visual inspection, which takes a lot of time and cognitive effort, this article aims to assist in this task to automate the diagnosis and make it more interpretable for the analyst. Using a public dataset, an architecture using the ResNet50 is built. Then, LIME and DT are invoked to analyze the interpretability of the model.

Both LIME and DT presented good performance regarding fidelity. However, the stability of LIME was lower than that of DT in some test examples. Through a visual analysis of the main parameters that each model used to explain the diagnoses, it was possible to perceive better stability of DT to LIME – since the features were always maintained in the test examples.

This work is part of a context little explored in the literature, up to now, since much of the research in the XAI area is focused on structured data and with the classic ML models. An example of this is that most existing XAI models are easily applicable post-hoc to ML models, which is not the case for DL methods. At the same time that XAI moves to meet existing (and necessary) regulations, it needs to be accessible to methods in the field of computer vision, as these are also widely used in the literature and applied to a multitude of problems.

ACKNOWLEDGMENTS

S. S. Santos would like to thank the Federal Center for Technological Education of Minas Gerais (CEFET-MG). The authors declare that this work has been supported by the Brazilian agencies CAPES, CNPq and FAPEMIG. MINDS Lab: `https://minds.eng.ufmg.br/`

REFERENCES

Alves, M.A., Castro, G.Z., Oliveira, B.A.S., Ferreira, L.A., Ramírez, J.A., Silva, R., and Guimarães, F.G. (2021). Explaining machine learning based diagnosis of COVID-19 from routine blood tests with decision trees and criteria graphs. *Computers in Biology and Medicine*, 132, 104335. doi:10.1016/j.compbiomed.2021.104335.

Brinker, T.J., Hekler, A., Enk, A.H., Klode, J., Hauschild, A., Berking, C., Schilling, B., Haferkamp, S., Schadendorf, D., Fröhling, S., et al. (2019). A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer*, 111, 148–154. doi:10.1016/j.ejca.2019.02.005.

Daghrir, J., Tlig, L., Bouchouicha, M., and Sayadi, M. (2020). Melanoma skin cancer detection using deep learning and classical machine learning techniques: A hybrid approach. In *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 1–5. IEEE. doi:10.1109/ATSIP49331.2020.9231544.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639), 115–118. doi:10.1038/nature21056.

Ferreira, L.A., Guimarães, F.G., and Silva, R. (2020). Applying genetic programming to improve interpretability in machine learning models. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, 1–8. IEEE. doi:10.1109/CEC48606.2020.9185620.

Fulton, L.V., Dolezel, D., Harrop, J., Yan, Y., and Fulton, C.P. (2019). Classification of Alzheimer's disease with and without imagery using gradient boosted machines and ResNet-50. *Brain sciences*, 9(9), 212. doi:10.3390/brainsci9090212.

Gulli, A. and Pal, S. (2017). *Deep learning with Keras*. Packt Publishing Ltd.

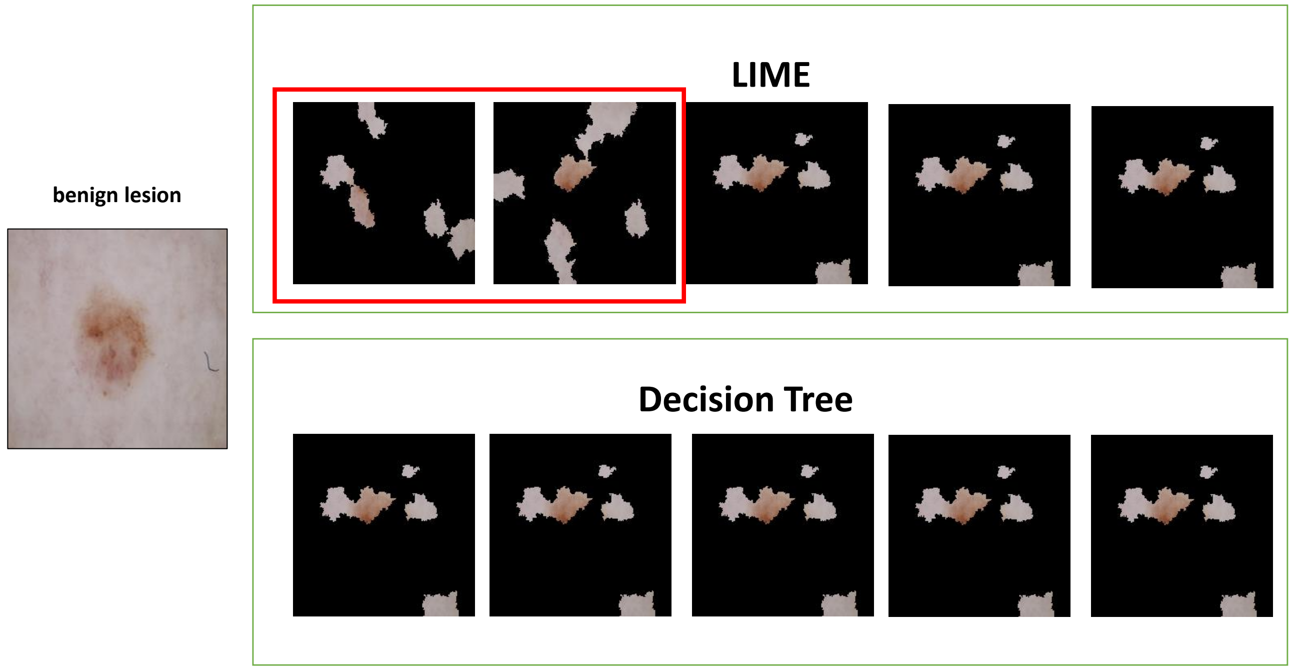


Figure 4. Explanations provided by LIME and Decision Tree for benign class.

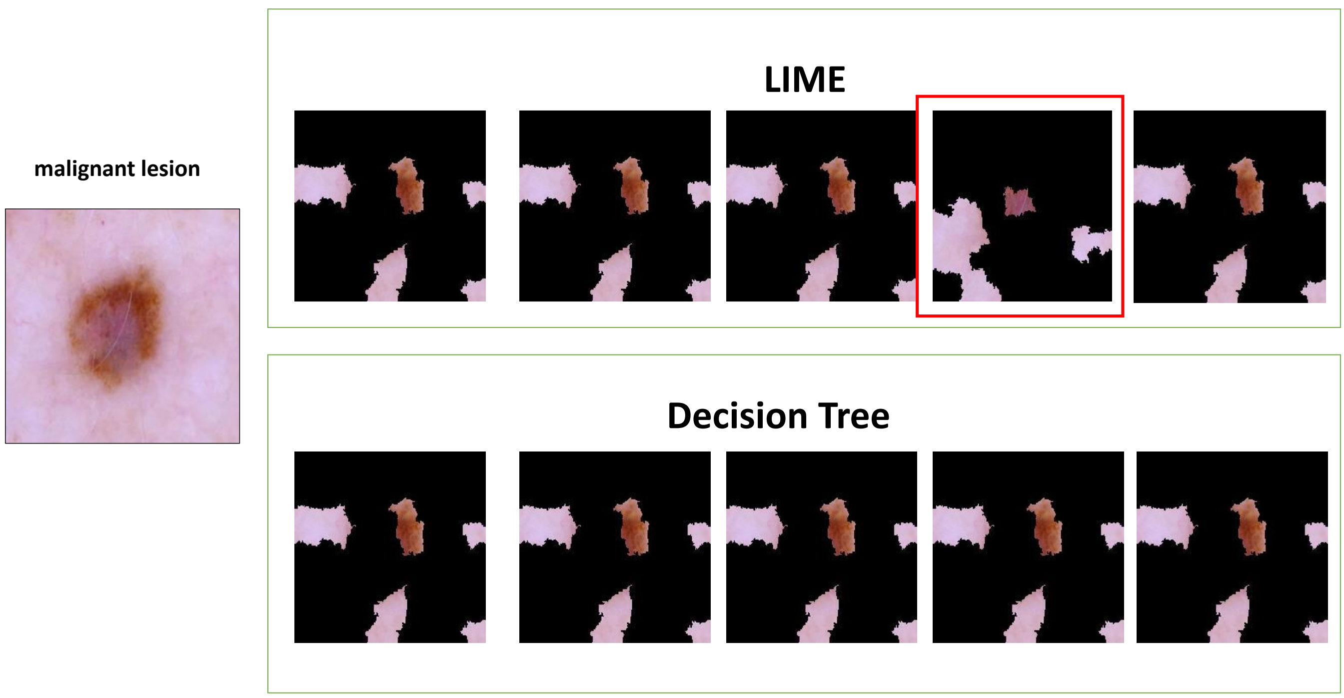


Figure 5. Explanations provided by LIME and Decision Tree for malignant class.

Hanif, A.M., Beqiri, S., Keane, P.A., and Campbell, J.P. (2021). Applications of interpretability in deep learning models for ophthalmology. *Current opinion in ophthalmology*, 32(5), 452–458. doi:10.1097/ICU.0000000000000780.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern* *recognition*, 770–778.

Hosny, K.M., Kassem, M.A., and Foaud, M.M. (2018). Skin cancer classification using deep learning and transfer learning. In *2018 9th Cairo international biomedical engineering conference (CIBEC)*, 90–93. IEEE. doi:10.1109/CIBEC.2018.8641762.

Jaleel, J.A., Salim, S., and Aswin, R. (2013). Computer aided detection of skin cancer. In *2013 International Conference on Circuits, Power and Computing Technologies (ICCPCT)*, 1137–1142. IEEE. doi:10.1109/ICCPCT.2013.6528879.

Javaid, A., Sadiq, M., and Akram, F. (2021). Skin Cancer Classification Using Image Processing and Machine Learning. In *2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST)*, 439–444. IEEE. doi:10.1109/IBCAST51254.2021.9393198.

Jiang, S., Li, H., and Jin, Z. (2021). A visually interpretable deep learning framework for histopathological Image-based skin cancer diagnosis. *IEEE Journal of Biomedical and Health Informatics*, 25(5), 1483–1494. doi:10.1109/JBHI.2021.3052044.

Kaggle (2019). Skin cancer: Malignant vs. benign. `https://www.kaggle.com/fanconic/skin-cancer-malignant-vs-benign`.

Loh, W.Y. (2011). Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1), 14–23. doi:10.1002/widm.8.

Magesh, P.R., Myloth, R.D., and Tom, R.J. (2020). An Explainable Machine Learning Model for Early Detection of Parkinson's Disease using LIME on DaTSCAN Imagery. *Computers in Biology and Medicine*, 126, 104041. doi:10.1016/j.compbiomed.2020.104041.

Mehta, P. and Shah, B. (2016). Review on techniques and steps of computer aided skin cancer diagnosis. *Procedia Computer Science*, 85, 309–316. doi:10.1016/j.procs.2016.05.238.

Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144. doi:10.1145/2939672.2939778.

Sachan, A. (2019). Detailed guide to understand and implement ResNets. *Learn Machine Learning, AI & Computer vision.*

Santos, S.S., Alves, M.A., Ferreira, L.A., and Guimaraes, F.G. (2021). PDTX: A Novel Local Explainer Based on the Perceptron Decision Tree. In *Anais do XV Congresso Brasileiro de Inteligência Computacional*. SBIC, Joinville, SC.

Thomas, S.M., Lefevre, J.G., Baxter, G., and Hamilton, N.A. (2021). Interpretable deep learning systems for multi-class segmentation and classification of non-melanoma skin cancer. *Medical Image Analysis*, 68, 101915. doi:10.1016/j.media.2020.101915.

Thurnhofer-Hemsi, K. and Dominguez, E. (2020). A convolutional neural network framework for accurate skin cancer detection. *Neural Processing Letters*, 1–21. doi:10.1007/s11063-020-10364-y.

Tschandl, P., Codella, N., Akay, B.N., Argenziano, G., Braun, R.P., Cabo, H., Gutman, D., Halpern, A., Helba, B., Hofmann-Wellenhof, R., et al. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *The Lancet Oncology*, 20(7), 938–947. doi:10.1016/S1470-2045(19)30333-X.

Van Der Velden, B.H., Kuijf, H.J., Gilhuijs, K.G., and Viergever, M.A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, 102470. doi:10.1016/j.media.2022.102470.

Van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., and Yu, T. (2014). scikit-image: image processing in Python. *PeerJ*, 2, e453. doi:10.7717/peerj.453.

Vidya, M. and Karki, M.V. (2020). Skin cancer detection using machine learning techniques. In *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 1–5. IEEE. doi:10.1109/CONECCT50063.2020.9198489.

Xiang, A. and Wang, F. (2019). Towards interpretable skin lesion classification with deep learning models. In *AMIA annual symposium proceedings*, volume 2019, 1246. American Medical Informatics Association.

Xie, P., Zuo, K., Zhang, Y., Li, F., Yin, M., and Lu, K. (2019). Interpretable classification from skin cancer histology slides using deep learning: A retrospective multicenter study. *arXiv preprint arXiv:1904.06156*.

Yu, J. (2018). The Art of Machine Learning: Looking deeper with LIME. URL `https://jyu-theartofml.github.io/posts/lime`.

Zhang, N., Cai, Y.X., Wang, Y.Y., Tian, Y.T., Wang, X.L., and Badami, B. (2020). Skin cancer diagnosis based on optimized convolutional neural network. *Artificial intelligence in medicine*, 102, 101756. doi:10.1016/j.artmed.2019.101756.