On the Use of Explainable Artificial Intelligence for the Differential Diagnosis of Pigmented Skin Lesions

Sandro Hurtado[1,2,3 [0000−0003−0990−480X]], Hossein Nematzadeh[1,2,4 [0000−0002−6161−0430]], José García-Nieto[1,2,3 [0000−0003−2985−3480]], Miguel-Ángel Berciano-Guerrero[3,5 [0000−0002−5437−5196]], and Ismael Navas-Delgado[1,2,3 [0000−0001−7819−5416]]

[1] Dept. de Lenguajes y Ciencias de la Computación, Universidad de Málaga, Spain
[2] ITIS Software, Arquitecto Francisco Peñalosa 18, 29071, Málaga, Spain
[3] Biomedical Research Institute of Málaga (IBIMA), Spain
[4] Dept. of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran
[5] Medical Oncology Intercenter Unit. Regional and Virgen de la Victoria University Hospitals, Málaga
sandrohr@uma.es

Abstract. In the last few years, eXplainable Artificial Intelligence (XAI) has been attracting attention in data analytics, as it shows great potential in interpreting the results of complex machine learning models in the application of medical problems. The nutshell is that the outcome of the machine learning-based applications should be understood by end users, specially in medical data context where decisions have to be carefully taken. As such, many efforts have been carried out to explain the outcome of a deep learning complex model in processes where image recognition and classification are involved, as in the case of Melanoma cancer. This paper represents a first attempt (to the best of our knowledge) to experimentally and technically investigate the explainability of modern XAI methods Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive exPlanations (SHAP), in terms of reproducibility of results and execution time on a Melanoma image classification data set. This paper shows that XAI methods provide advantages on model result interpretation in Melanoma image classification. Concretely, LIME performs better than SHAP gradient explainer in terms of reproducibility and execution time.

Keywords: eXplainable Artificial Intelligence · Melanoma medical image classification · LIME · SHAP · Deep Learning

1 Introduction

Explainable AI (XAI) is an artificial intelligence approach oriented to explain the results of complex machine learning algorithms [2]. Generally, it is believed that

as the complexity of a machine learning algorithm increases, the understandability of the results become harder [3]. Previously, the robustness of a classification algorithm was evaluated using well-known criteria such as accuracy, precision, recall, Fscore, and etc. However, in real-world scenarios, human experts usually prefer the use of understandable algorithms, even though they usually have moderate, sometimes limited, performance that other complex black-box techniques, such as deep learners. In fact, explainability besides accuracy are two important factors to assess the output of any machine learning algorithms [9]. One of the main categories of explainers are post hoc model-agnostic. Post hoc refers to those methods that are applied after training the model and not at the middle of the model training process. Model-agnostic refers to the group of explainers that are not specifically designed for a certain machine learning algorithm. XAI specifically well-adapted to provide explanation ability to deep learning output on medical datasets [5], where Melanoma cancer is not an exception.

Melanoma is the most aggressive skin tumour, with a 5-year survival rate of 93% if diagnosed in early stages, but only 27% if diagnosed at an advanced stage with the presence of metastatic disease[6]. In Spain, 6,108 cases of melanoma were estimated in 2021 (2,480 men and 3,678 women), being the fifth most frequent cancer in men and women [7]. Diagnosis in the early stages is what allows better survival rates, although it entails the difficulty of differentiating it from other pigmented skin lesions (nevus and seborrheic keratosis, mainly), which are followed up. The inclusion of artificial intelligence in the diagnosis would allow a more accurate diagnosis. In concrete, there are many efforts to melanoma diagnosis using deep learning [1][7]. In order to realize trustworthy AI, XAI can be used as a technical method to ensure transparency of deep learning by helping better understand the neural network's underlying mechanisms and explaining system behaviours to users (in our case clinicians).

This paper is, to the best of our knowledge, a first attempt to evaluate two well-known post hoc model-agnostic methods in XAI, namely: Local Interpretable Model-Agnostic Explanations (LIME) [8] and SHapley Additive exPlanations (SHAP) [6], on explaining the deep learning prediction on Melanoma image dataset technically. Reproducibility and execution time are introduced as two major criteria for comparing LIME and SHAP. This paper finally concludes which of the aforementioned method is most suitable for explanation of Melanoma detection from an engineering point of view. The rest of this paper is organized as follow. Section 2 provides related information for LIME and SHAP. Section 3 demonstrates the methodology and the results achieved. Finally, section 4 concludes the paper by summarizing the findings.

[6] Melanoma Cancer statistics approved by the Cancer.Net Editorial Board, 01/2021 https://www.cancer.net/cancer-types/melanoma/statistics

[7] https://seom.org/images/Cifras_del_cancer_en_Espnaha_2021.pdf

2 Preliminaries

This research focuses on the model-agnostic AI explainers, which provide post-hoc interpretability i.e. why the prediction model predicted its output through providing after-the-fact evidence for the outputs. These explainers are probably the most popular ones in the current literature, which consist in Local Interpretable Model-Agnnostic Explanations (LIME) [8] and SHapley Additive exPlanations (SHAP), both comprising a group of techniques that help humans visualize what an already-trained model thinks is important.

LIME uses Equation 1 to minimize $\xi(x)$ so that f is the prediction model which is assumed as black box, g is a model in G as a class of potentially interpretable models that tries to approximate f, Π_x is used to define the locality around the sample to be explained (perturbations from x), and $\Omega(g)$ represents the complexity of explanation that should be minimized as well as $L(f, g, \Pi_x)$.

$$\xi(x) = argmin_{g \in G}\ L(f, g, \Pi_x) + \Omega(g) \tag{1}$$

SHAP values are concepts coming from game theory [6]. Shapely quantifies to what extent each player (features) contributes to the game (output of prediction model). Shapely creates a power set of features firstly. The cardinality of power set is 2^n where n is the total number of features. Likewise, SHAP also requires to train 2^n models with different set of features according to the power set. It is obvious that as the number of features is higher the number of models to be trained increases exponentially, which is treated by Lundberg et al.[6] through some approximations and samplings. Basically, calculating SHAP values has two steps, namely calculating marginal contributions of each feature and weighing the marginal contributions which can be shown in general in Equation 2, so that F is the entire number of (f) features and $set = 1, ..., F$.

$$SHAP_f(x) = \sum_{f \in set} [|set| \times \binom{F}{|set|}]^{-}1[Predict_{set}(x) - Predict_{set/f}(x)] \tag{2}$$

Fig. 1 illustrates the difference between SHAP and LIME in general. According to this figure, LIME initially perturbs the sample to explain x to create the set $Z = z_1, z_2, ..., z_m$. Next, it selects an interpretable model (such as linear regression) to calculate the importance of features (calculating the coefficients related to each feature) via $g(Z)$. LIME finally selects the most effective features (through sorting coefficients if g is linear regression). However, SHAP build SHAP values by calculating the marginal contribution of features and weighing them. Effective features are those with greater SHAP values. Moreover, summing the SHAP values gives exactly the difference between the output of full model and null model, which shows the additive explanations of SHAP.

While SHAP explainers are model agnostic, there exists two variations that could be used for deep learning, namely deep explainer and gradient explainer. Deep explainer approximates the conditional expectations of SHAP values using a selection of background samples, while gradient explainer explains a model

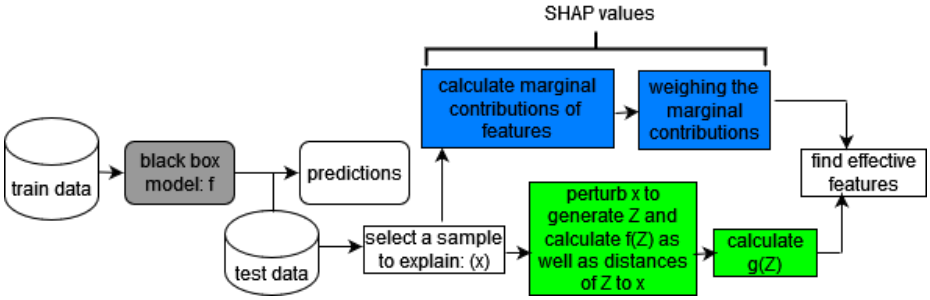

Fig. 1. General workflow of SHAP and LIME

using expected gradients which reformulates the integral as an expectation and combines that expectation with sampling reference values from the background dataset.

3 Methodology

The methodology of the paper is illustrated using a pipeline in Fig. 2. The image dataset is online available in Kaggle Skin Lesion Images for Melanoma Classification (ISIC2019) repository[8]. It comprises more than 25,000 images with imbalanced classes (the majority of training data is nevus) which could cause an erroneous accuracy and incorrect predictions. There are many methods to balance training data including undersampling the majority class, oversampling the minority classes, applying SMOTE, and etc depending on the dataset. However, our experiments reveal that the best technique for image datasets like Melanoma is the combination of random oversampling the minority classes following by applying data augmentation.

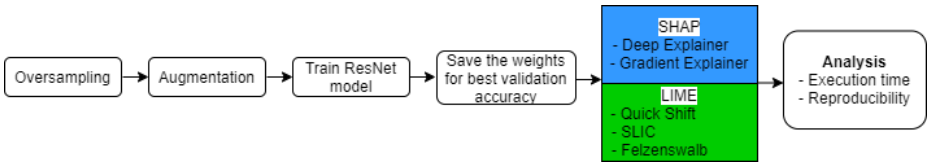

Fig. 2. Proposed methodology.

Thus, in the preprocessing step, the distribution of classes were equalized using random oversampling initially. Oversampling solely can lead to overoptimism in prediction. Assuming the training data is split into train and validation sets. It is expected that some images in the training data appear in the validation

[8] In URL https://www.kaggle.com/andrewmvd/isic-2019

set, since there exist multiple replicated images as a result of random oversampling the minority classes. As such overfitting could happen where the model prediction will be high in training data, but very low in unseen data. Here data augmentation could alleviate overfitting. The data augmentation in this study is done through rescaling, rotating, width-shift, height-shift, and horizontal-flip augmentation. The pipeline in Fig. 2 follows by applying pre-trained ResNet [4] convolutional Deep Learning model and saving the best weights. Then, model agnostic post hoc explainers (SHAP with Deep and Gradient explainers, LIME with three well-known segmentation algorithms) are used to evaluate the results based on reproducibility of the results and execution time.

Reproducibility means the ability of the method to successfully reproduce same explanations in multiple runs. Likewise, execution time refers to the elapsed time starting from creating the explainer until calculating the explanation and generating the pictorial results. Table 1 also shows the main characteristics of the Melanoma dataset prior to oversampling and augmentation. After oversampling the distribution of each class in training set is equal to 1,372 so that the the entire training set contains 4,116 observations.

Table 1. Melanoma dataset description after oversampling class imbalance

Data	#Observations	Distribution of observations
Train	2000	374/Melanoma, 1372/Nevus, 254/Seborrheic-keratosis
Validation	150	30/Melanoma, 78/Nevus, 42/Seborrheic-keratosis
Test	600	117/Melanoma, 393/Nevus, 90/Seborrheic-keratosis

Table 2. Description of four selected samples for experimentation

Test observation	Real label	Melanoma	Nevus	Seborrheic-keratosis
Sample 1	Nevus	0.31	0.57	0.12
Sample 2	Melanoma	1.00	0.00	0.00
Sample 3	Nevus	0.00	1.00	0.00
Sample 4	Seborrheic-keratosis	0.00	0.00	1.00

3.1 Evaluation

This section provides related information for calculated metrics. All the experiments have been conducted in a virtualization environment running on a private high-performance cluster computing platform. This infrastructure is located at the Ada Byron Research Center at the University of Málaga (Spain), and comprises a number of IBM hosting racks for storage, units of virtualization, server compounds and backup services. Our virtualization platform is hosted in this

computational environment. Concretely, this platform is made up of a CPU with Intel(R) Xeon(R) Gold 6130 @ 2.10GHz, maximum 2 TB of HDD, maximum 64 GB of RAM, and Ubuntu 20.04.3 LTS(GNU/Linux 5.4.0-1049-kvm x86_64).

Since it is impossible to illustrate the entire test samples four test samples were selected to investigate the reproducibility and execution time analysis as explained in Table 2, so that for each sample the real labeling and the prediction of deep learning are shown.

3.2 Evaluation of LIME

Fig 3 illustrates the reproducibility of LIME using three well-known segmentation algorithm namely, quick shift, Simple Linear Iterative Clustering (SLIC) and felzenswalb. Quick shift uses approximation of kernelized mean-shift and it belongs to the family of local mode-seeking algorithms. SLIC uses k-means which is a simpler clustering method in comparison with the clustering method in quick shift. In contrast, felzenswalb uses a graph-based approach for image segmentation.

Fig 3 is the result of 5 multiple runs of LIME algorithm for 5 top features with different number of perturbations regarding each of the four images in Table 2. The original segmentation is illustrated for each image using quick shift, SLIC, and felzenswalb in Fig 3 initially, so that the segmentation algorithms are tuned to contain roughly same number of segments for each algorithm, to have a fair comparison between them. Under each image is a fraction that shows how many times LIME is able to generate exactly same 5 top features in 5 multiple runs using each segmentation algorithm. For example, 4/5 for sample 1 with quick shift algorithm and 5,000 perturbations means the result of LIME in 4 runs from 5 runs are exactly same. As such, sample 1 achieves 1/5 for 100 perturbations using quick shift algorithm, which means that there are five unique results so that one of them is selected randomly.

It is noteworthy mentioning that, quick shift and SLIC have relatively same segmentation trend in comparison with felzenswalb so that this last sometimes results in segments with sizes that vary greatly as in sample 2 and sample 3. This may affect the reproducibility of LIME either positively in sample 3 or negatively in sample 2.

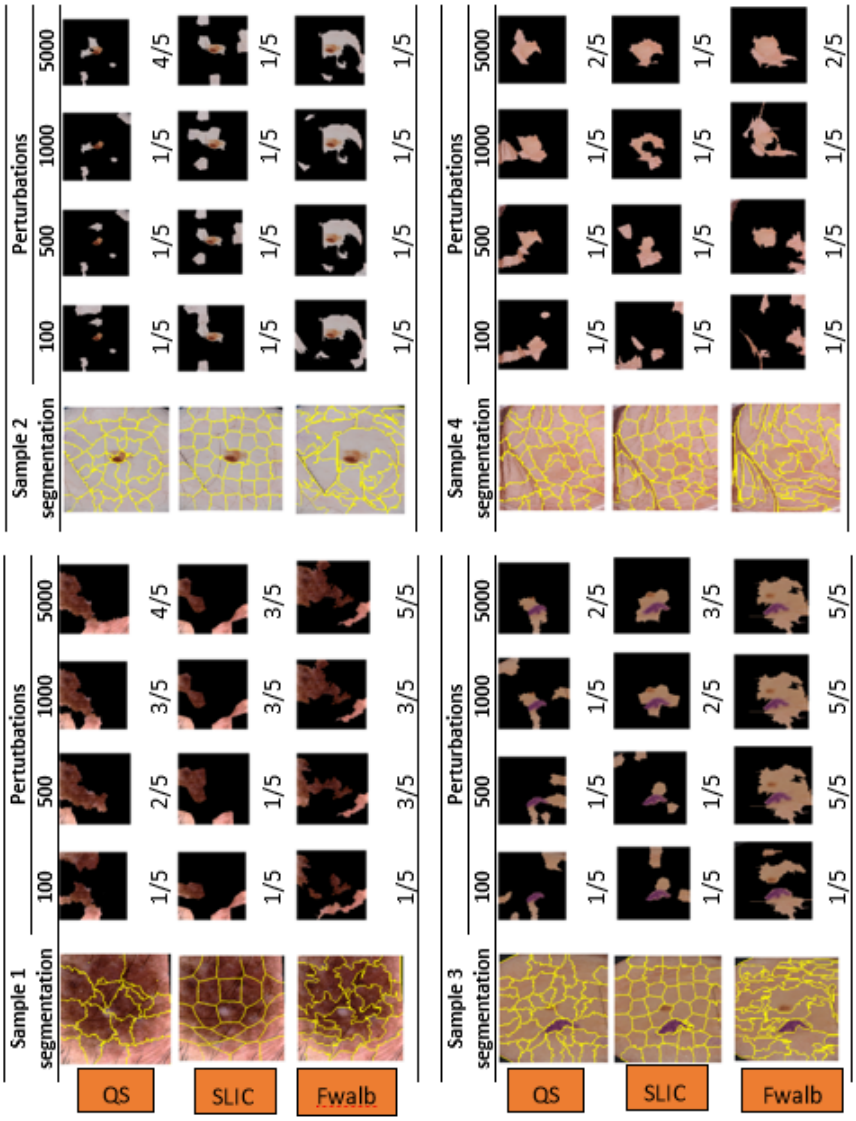

Fig. 3. Strict analysis of LIME reproducibility by increasing number of perturbations

While Fig 3 analyzes reproducibility strictly, Fig 4 checks the reproducibility of LIME more gently by calculating the number of features in each perturbation (100, 500, and 1,000) that have also been observed when the perturbation is 5,000. Fig 4 shows that as the number of perturbation increases from 100 to 1,000, more features from that perturbation are observed within 5,000 perturbation. If two superpixels are equally good at explaining, LIME may pick an arbitrary one which sometimes result in not reproducible explanations. Fig 4 shows that by increasing the number of perturbation, LIME converges to reproducibility.

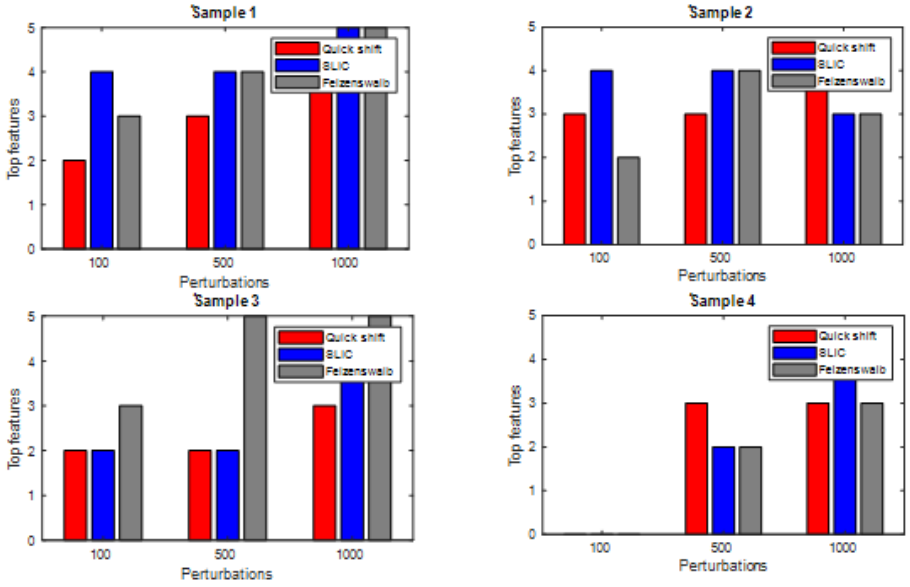

Fig. 4. Gentle analysis of LIME reproducibility by increasing number of perturbations

Recalling that good segmentation often depends on the application, illustrations in Fig 3 and 4 show that the reproducibility in LIME mostly increases while the number of perturbation increases from 500 to 5,000 using any segmentation algorithm (the default number of perturbation in LIME is 1,000). While increasing number of perturbations has a positive effect in reproducibility of LIME, another approach is to fix the random seed to initialize the random number generator. This way, using any number of perturbations the explainability results are same. Nonetheless, greater number of perturbations together with fixed random seed result in better accuracy as well. Nonetheless, Fig 5 shows how successfully LIME recognizes regions contributed to target label by increasing the number of perturbations and using fixed random seed. This last figure also reveals that

LIME intelligently did not recognize mm scale and hair as effective features, but considers the stain in sample 3 within 5 most important superpixels.

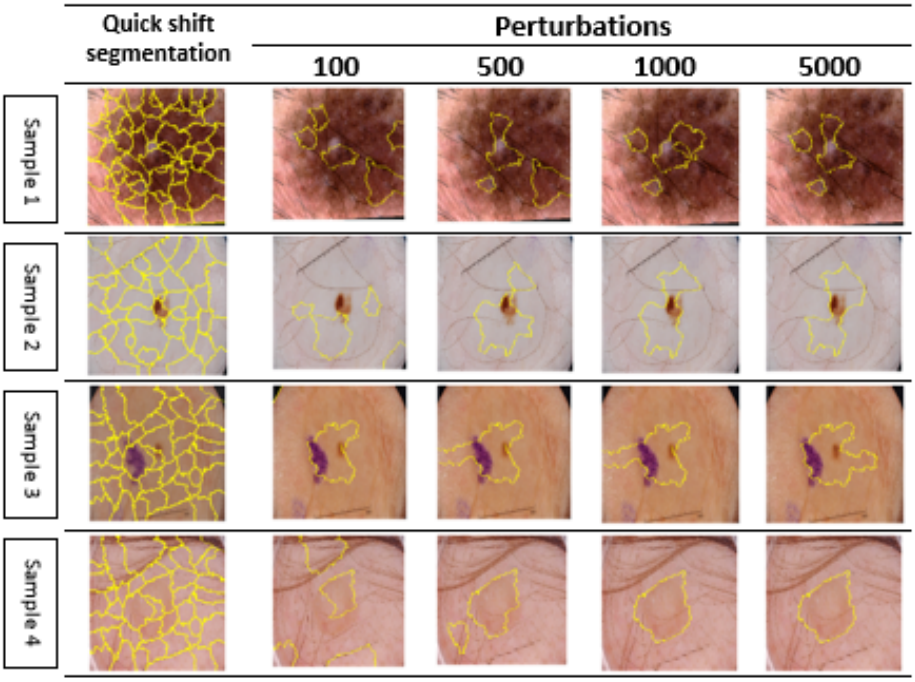

Fig. 5. Reproducibility analysis of LIME using fixed random seed and variable number of perturbations

3.3 Evaluation of SHAP

As commented before, there exists two variations of SHAP optimized for deep learning, namely gradient explainer and deep explainer. The SHAP kernel explainer could also be used because it works for all models, but is slower than the other model type-specific algorithms, as it makes no assumptions about the model type. Thus, to avoid redundancy of figures with same results and for the sake of hardware limitation (passing more than 100 background data was unreasonably expensive), the reproducibility of SHAP has been tested using solely with gradient explainer, shown in Fig 6. Generally, pink pixels contribute to the model output and blue pixels contribute not being of that class. The intensity of color shows the intensity of contribution. Since gradient and deep explainer explains the prediction using pixels and not superpixels, it is difficult to trace the reproducibility numerically as it was done for LIME.

The nsamples parameter in gradient explainer (by default = 200) indicates the number of samples are taken to compute the expectation and shows accuracy of explanation. This gives better estimates of SHAP values as the nsamples increases, which leads to low variate estimation of the SHAP values, however the execution time increases. Fig 6 shows that as the nsamples increases from 100 to 5,000 the explainability becomes a bit more reproducible, which is less obvious in sample 1 because the deep learning model is not completely sure about its prediction. Fig 6 also shows that gradient explainer considers the stain in sample 3 same as LIME in Fig 5. The gradient explainer in Fig 6 uses the entire 4,116 images in train set as a background data (the random seed in calculation of SHAP values is set to 42).

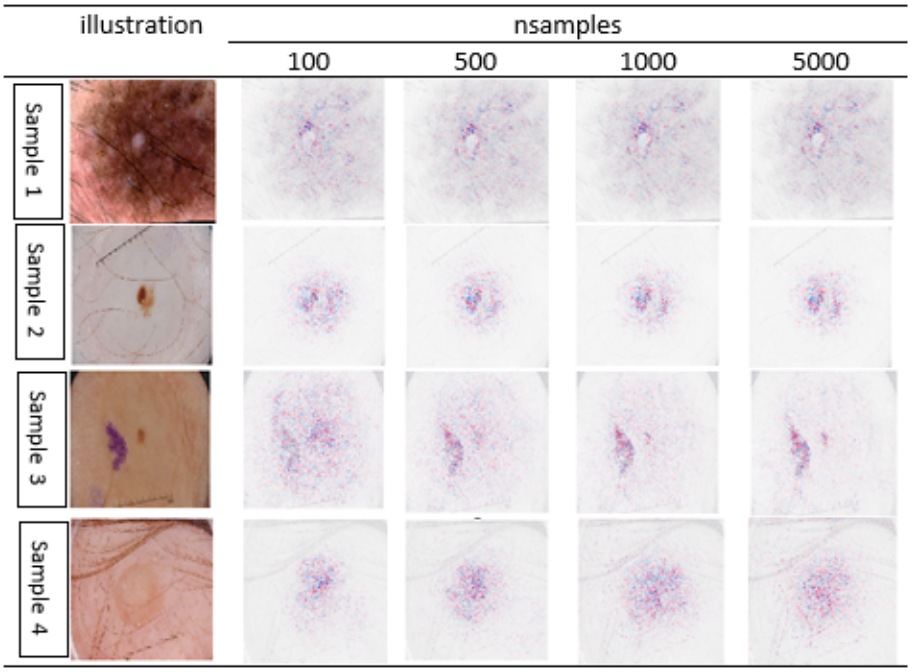

Fig. 6. Visual reproducibility analysis of SHAP gradient explainer

3.4 Computational Effort

From the point of view of the computational effort, Fig 7 compares LIME (using quick shift) and SHAP gradient explainer in terms of execution time, so that N is the number of perturbation and nsamples for LIME and SHAP, respectively. It is clear that LIME spends less amount of time for explainability as N increases, while SHAP gradient explainer is almost three times slower than LIME. It is

noteworthy mentioning that changing the segmentation algorithm does not have a considerable difference in execution time of LIME. SLIC is very competitively faster than quick shift and also quick shift is very closely faster than felzenswalb. Thus, Fig 7 the better performance in terms of execution time of LIME using quick shift as a moderate segmentation algorithm. Technically speaking, LIME has more reproducibility power and is almost much faster than SHAP gradient on melanoma dataset. Thus, there are sufficient engineering justifications to use LIME for explanability of deep learning on melanoma dataset for a single prediction rather than SHAP gradient explainer.

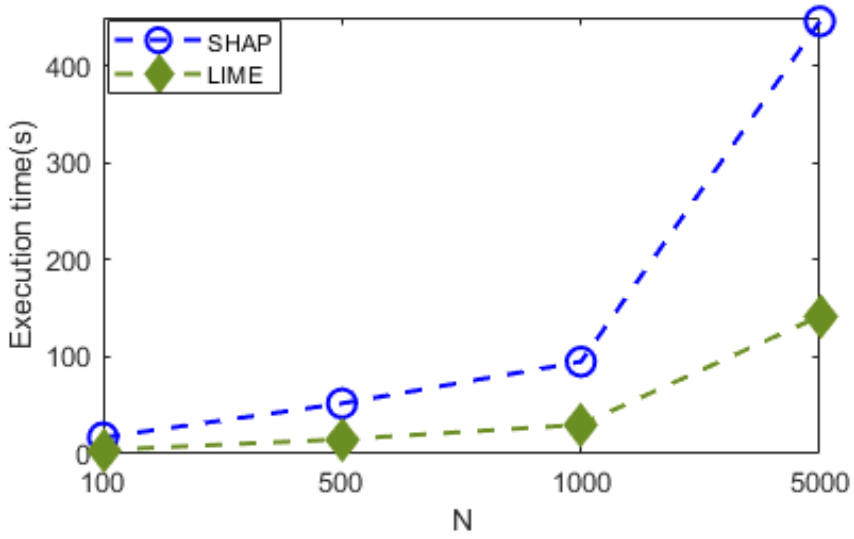

Fig. 7. Efficiency of LIME Vs SHAP

4 Conclusion

This paper investigated the explainability of Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) in order to help in the differential diagnosis of pigmented skin lesions. The evaluation criteria focuses on the reproducibility of the results, as well as the execution time. Three variations of LIME (using three well-known segmentation algorithms) are used and gradient explainer is selected for SHAP. From the engineering point of view, LIME was faster than SHAP. The idea is that while acceptable results are achieved by LIME in the case of differential diagnosis of pigmented skin lesions, there is no need to use SHAP because of its expensive efficiency. LIME works with super pixels and the reproducibility of results were more controllable than

SHAP gradient explainer. Thus, it can be concluded that XAI methods show potentials in providing interpretable results for the specific case of pigmented skin lesions classification, in the context of Melanoma cancer diagnosis. Specifically, LIME shows better performance than SHAP gradient explainer in terms of reproducibility and execution time.

The general idea for future work is to approach explainability of deep learning on melanoma data set through improving LIME, as well as to tacking with other different kind of medical image datasets.

Acknowledgement

This work has been partially funded by the Spanish Ministry of Science and Innovation via Grant PID2020-112540RB-C41 (AEI/FEDER, UE) and Andalusian PAIDI program with grant P18-RT-2799.

References

1. Banerjee, S., Singh, S.K., Chakraborty, A., Das, A., Bag, R.: Melanoma diagnosis using deep learning and fuzzy logic. Diagnostics **10**(8) (2020). https://doi.org/10.3390/diagnostics10080577, https://www.mdpi.com/2075-4418/10/8/577
2. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion **58**, 82–115 (2020). https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.012, https://www.sciencedirect.com/science/article/pii/S1566253519308103
3. Gulum, M.A., Trombley, C.M., Kantardzic, M.: A review of explainable deep learning cancer detection models in medical imaging. Applied Sciences **11**(10) (2021)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
5. Knapič, S., Malhi, A., Saluja, R., Främling, K.: Explainable artificial intelligence for human decision support system in the medical domain. Machine Learning and Knowledge Extraction **3**(3), 740–770 (2021)
6. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 4768–4777. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
7. Naeem, A., Farooq, M.S., Khelifi, A., Abid, A.: Malignant melanoma classification using deep learning: Datasets, performance measurements, challenges and opportunities. IEEE Access **8**, 110575–110597 (2020). https://doi.org/10.1109/ACCESS.2020.3001507
8. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. Association for Computing Machinery (2016)
9. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.: Toward interpretable machine learning: Transparent deep neural networks and beyond. CoRR **abs/2003.07631** (2020)