CRP and LRP on classifying medical images, including lung and skin lesions

Paweł Pawlik pp406289@students.mimuw.edu.pl
Michał Siennicki ms406340@students.mimuw.edu.pl
Alicja Ziarko az406665@students.mimuw.edu.pl
University of Warsaw, Poland

Abstract

The goal of our project was to examine two model explanation methods - Concept Relevance Propagation (CRP) (Achtibat et al., 2022) and Layer-Wise Relevance Propagation (LRP) (Montavon et al., 2019) on medical datasets - Melanoma (Rotemberg V., 2021) and COVID-QU-Ex (Tahir et al., 2021). We present a comparison of the results of applying those methods to two distinct models.

1. Introduction

In recent years deep learning models have achieved tremendous success in computer vision. One application of them is diagnosing medical conditions from photos and scans. In comparison to other domains, the possibility of a model being wrong is much more serious. Hence treating it as a black box is not a possibility. For that reason, it is essential to be able to make use of interpretability methods. In this paper, we apply two different explainability methods to investigate the predictions of two different models on two different medical datasets. We were able to verify which features are important to the models. Furthermore, we investigate how the explanations from the data points change when augmentations are applied to them.

Related Work Methods of explainable machine learning have been successfully applied to aid diagnostics. In (Wu et al., 2021) many methods for interpreting Covid-19 severity predictions were investigated and compared. In (Mahima et al., 2021) explainable AI is used to analyze the robustness of Neural Networks to adversarial examples. Best to our knowledge, there was no analysis of how augmentations affect explanations of neural network predictions on medical data.

2. Methodology

The models, datasets, augmentations, and training results are described in Appendix A.

LRP Layer-Wise Relevance Propagation (Montavon et al., 2019) is a method designed to explain CNN-based models. LRP uses the network weights and the neural activations created by the forward pass to propagate the output back through the network until the input layer.

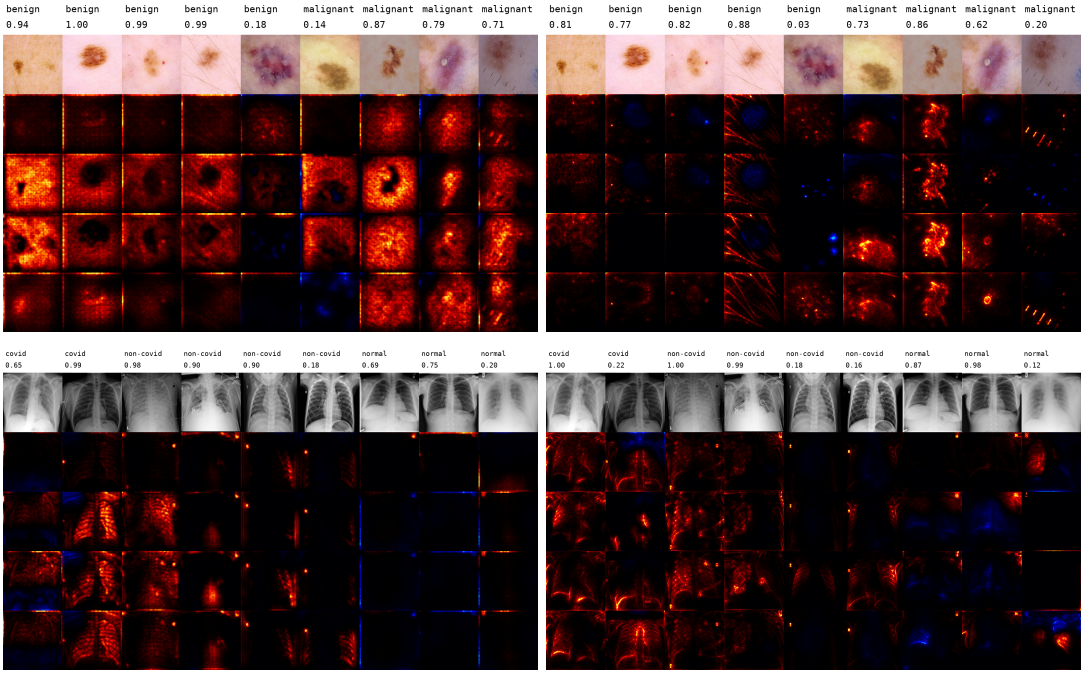

Figure 1: Melanoma LRP and CRP results (top left: ResNet50, top right: VGG16) and COVID LRP and CRP results (bottom left: ResNet50, bottom right: VGG16). On each image the rows are: class name; prediction on the class; input image; LRP explanations with composite EpsilonPlusFlat; 3 CRP explanations conditioned on different layers, from most shallow to deepest. More composites for LRP can be viewed in Appendix B

CRP Concept Relevance Propagation (Achtibat et al., 2022) can be seen as a generalization of the Layer-Wise Relevance Propagation method. The main difference is that it allows for propagating the relevance in a more controlled manner - arbitrary network elements can be masked. In this work, we are focusing on the global concept importance. We use this feature to check for each network layer which present concept is the most important. The exact setup is described in Appendix C.

3. Experimental results

We can observe LRP and CPR explanations for both architectures and datasets in Figure 1. In general, the LRP in CRP results are quite similar and can lead to related conclusions. The CRP method gives us more explanations and works on different layers which gives us better insights into the models' prediction understanding.

Melanoma Dataset We can observe that both models make strong use of features from the skin lesions, the measuring marks, hair, and the border of the image. Using the border of the image makes sense as in the Melanoma dataset we can observe that benign samples tend to have a light pink background (skin), while malignant are typically darker around

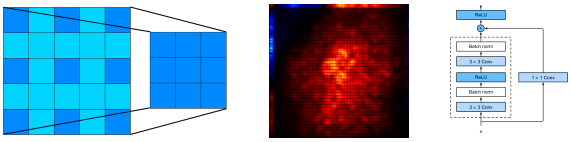

Figure 2: 1x1 convolution layer with stride 2 (left), zoomed ResNet50 LRP explanation (middle), ResNet block (right).

the lesion. Focusing on the measuring marks is probably undesired and may indicate some bias in the dataset. The presence of hair might be correlated with gender and men are at a higher risk of developing melanoma (Rastrelli et al., 2014), so it makes sense.

COVID Dataset The models focus on three main aspects: left/right X-ray lead markers, the overall lung area, and the border of the image.

The roentgen left/right markers are used by radiologists to label which lung is which one and are visible as single letters on the images. The explanations suggest that both methods find those letters important. This is highly undesired - a possible explanation for this bias is that our dataset could be unequally distributed between clinics or laboratories that use different markers.

The overall lung area is where we expect the model to focus. Here the VGG16 explanations again look more convincing. The most interesting cases are the ones where we can see both negative and positive attribution from different parts of the image. For example, in the right-most column, we can see a wrongly classified healthy lung (0.12 prediction from this class). The left lung (the right one in real life according to the marker) looks healthy and has a positive attribution. The other one is smaller, less visible, and has a negative impact on the score, which is reasonable.

The border of the image is the least desired feature here. The ResNet50 model focuses quite intensely on it. It might mean that there are some more biases in the data and perhaps some other augmentations should be used.

Additionally, the explanations for the 'normal' lungs are the least visible. This might make sense because of our asymmetrical dataset - there might be no characteristics of healthy lungs, just some features of sick ones.

Checker pattern There is one worrisome characteristic visible in the ResNet explanations - a visible checker pattern. This phenomenon is caused by the residual connections present in the ResNet architecture (Figure 2). Residual blocks that scale down the image use a 1x1 convolution with a stride equal to 2. This results in the checker pattern visible in LRP and CRP explanations as the signal is much stronger in the pixels that were used in residual connections.

ResNet vs VGG The LRP results are better for the VGG architecture on both datasets - the VGG manages to better highlight lung or skin changes shapes. Additionally, VGG's explanations have more often parts with both positive and negative attribution on the same image, which is helpful in interpreting those results. The differences come from the architectures - ResNet cannot be naturally divided into layers as opposed to VGG with

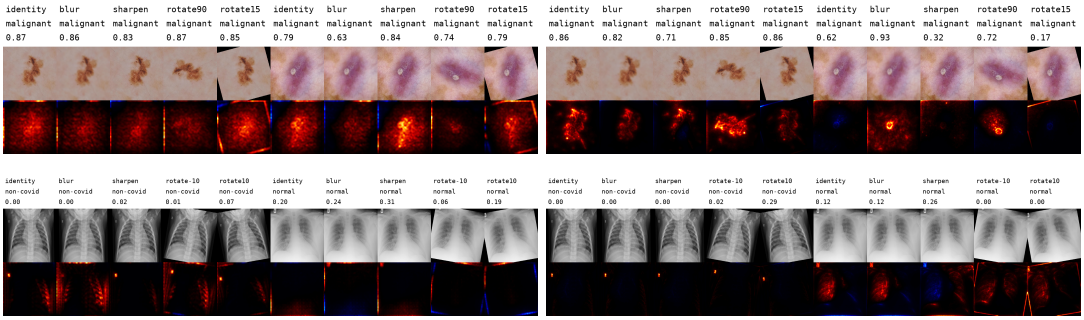

Figure 3: Results for LRP after input augmentation. Melanoma LRP results (top left: ResNet50, top right: VGG16) and COVID LRP results (bottom left: ResNet50, bottom right: VGG16). On each image, the rows are augmentation name, class name, prediction on the class, input image, and LRP explanations with composite EpsilonPlusFlat. More composites can be viewed in Appendix B

no residual blocks. In CRP both models give different, interesting results. In Melanoma ResNet does not extract the hair and measuring marks, which might be better. The borders of skin lesions are more clearly visible. In the COVID dataset, ResNet and VGG look at quite different features. Specialist knowledge is required to verify which features are more important.

Robustness To evaluate the robustness of the LRP we compared the visualizations after applying different augmentations to the input image. We have tested blur, sharpen, and rotations (Figure 3). We can observe that the explanations change drastically in the case of the VGG architecture. On the COVID dataset, the positive and negative heatmap can even swap places. On the Melanoma dataset, we can observe that the same skin lesion once attributed positively and once negatively depending on augmentations. ResNet results are more consistent, although they are still quite different. We can conclude that LRP evaluations for CNNs are not robust. Most likely, this is due to the known CNNs' vulnerability to augmentations and adversarial attacks (Madry et al., 2017).

4. Conclusion

There are some interesting features that are used by the networks for predictions in both datasets and models. There seems to be some bias in the data, but there are some concepts that are possible for a human to interpret. CRP and LRP can therefore be used for explaining medical data. As shown throughout our report, when doing that one needs to be aware of the model structure influencing the relevance flow and also the lack of robustness to augmentations. Moreover, the features used for classification differ by model. Looking at the images and attributions from CRP would most definitely be a fascinating read for a radiologist that could firstly allow them to discover something they didn't previously see as a predictor of an illness and secondly allow us to be sure as to whether the model is using features that might actually have meaning.

References

Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From "where" to "what": Towards human-understandable explanations through concept relevance propagation, 2022. URL https://arxiv.org/abs/2206.03208.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2017. URL https://arxiv.org/abs/1706.06083.

KT Yasas Mahima, Mohamed Ayoob, and Guhanathan Poravi. An assessment of robustness for adversarial attacks and physical distortions on image classification using explainable ai. In *AI-Cybersec@ SGAI*, pages 14–28, 2021.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. *Layer-Wise Relevance Propagation: An Overview*, pages 193–209. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28954-6. doi: 10.1007/978-3-030-28954-6_10. URL https://doi.org/10.1007/978-3-030-28954-6_10.

Marco Rastrelli, Saveria Tropea, Carlo Riccardo Rossi, and Mauro Alaibac. Melanoma: epidemiology, risk factors, pathogenesis, diagnosis and classification. *In vivo*, 28(6):1005–1011, 2014.

Betz-Stablein B. Rotemberg V., Kurtansky N. A patient-centric dataset of images and metadata for identifying melanomas using clinical context., 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. URL https://arxiv.org/abs/1409.1556.

Anas M. Tahir, Muhammad E.H. Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M. Sohel Rahman, Somaya Al-Maadeed, Sakib Mahmud, Maymouna Ezeddin, Khaled Hameed, and Tahir Hamid. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in Biology and Medicine*, 139:105002, 2021. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2021.105002. URL https://www.sciencedirect.com/science/article/pii/S0010482521007964.

Han Wu, Wenjie Ruan, Jiangtao Wang, Dingchang Zheng, Bei Liu, Yayuan Geng, Xiangfei Chai, Jian Chen, Kunwei Li, p Shaolin Li, and Sumi Helal. Interpretable machine learning for covid-19: An empirical study on severity prediction task. *IEEE Transactions on Artificial Intelligence*, pages 1–1, 2021. doi: 10.1109/TAI.2021.3092698.

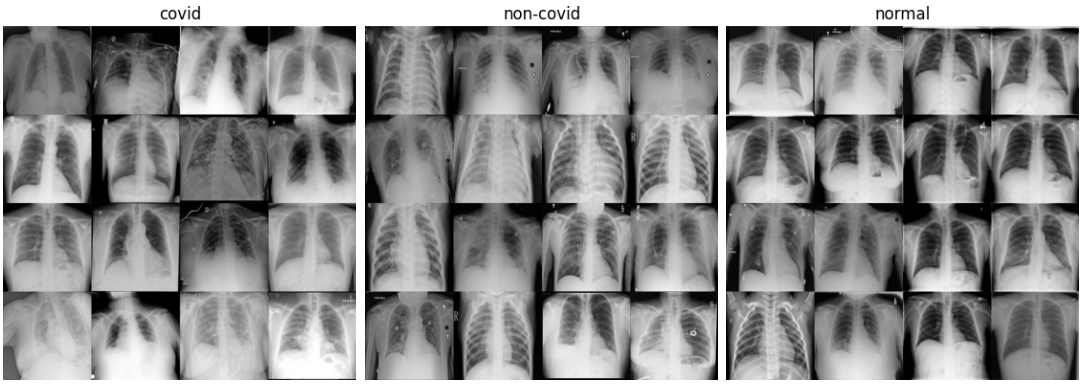

Figure 4: COVID dataset samples divided by class.

Appendix A. Models, Datasets

A.1 Models

We have tested two different Convolutional Neural Network (CNN) architectures - ResNet50 (He et al., 2015) and VGG16 (Simonyan and Zisserman, 2014). VGG16 is built as a series of convolutional and pooling layers with a dense head. ResNet50 on the other hand has residual connections, which counters the fading gradients problem. This fact is important to understand some differences in interpretation visualizations between those two architectures.

A.2 COVID-QU-Ex

Description The COVID-QU-Ex dataset (Tahir et al., 2021) consists of 33,920 chest X-ray (CXR) images including 11,956 COVID-19 images, 11,263 non-COVID infections (Viral or Bacterial Pneumonia) and 10,701 healthy chest X-rays. It can be found at `https://www.kaggle.com/datasets/anasmohammedtahir/covidqu`. Samples can be seen in Figure 4.

Training 20% of the dataset was used as the test set (as proposed by the dataset authors). Using the VGG16 we obtained 98% ROC AUC on the test set (mean from the AUC for those three classes). With ResNet50 we obtained 94% ROC AUC.

A.3 Melanoma

Description The Melanoma dataset (Rotemberg V., 2021) consists of 37,648 dermoscopic training images of unique benign and malignant skin lesions (5,106 malignant and 32,542 benign samples). It can be found at `https://www.kaggle.com/datasets/nroman/melanoma-external-malignant-256`. Samples can be seen in Figure 5.

Augmentations The first trained networks were mostly focusing on the edges of the image. It turns out that some images from microscopes are circle shaped and have a black background outside the circle. Furthermore, this property is not balanced between the positives and negatives. We have decided to include an augmentation that simulates that black background from microscopes, which resolved this bias. The final augmentations

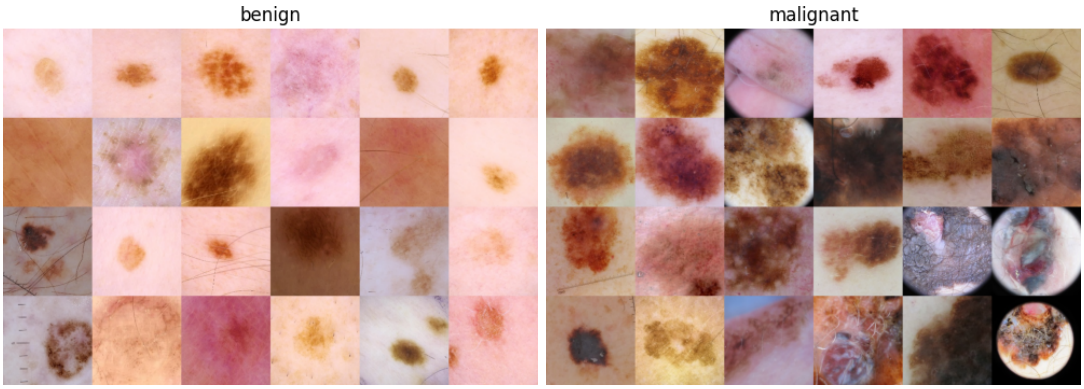

Figure 5: Melanoma dataset samples divided by class.

consist of Color Jitter, Random Perspective, Random Erasing, Random Rotation, and Microscope Simulation (as described above).

Training 33% of the dataset was used as the test set. Using the VGG16 we have managed to obtain 94% ROC AUC on the test set. Using ResNet50 we obtained 93% ROC AUC.

Appendix B. Enhanced LRP visualizations

In this appendix, we present results for the LRP method with different composites (Figure 6 and Figure 7). We found out that EpsilonPlusFlat works best so in the main report we only focus on the mentioned composite.

Appendix C. CRP experimental setup

The setup we followed for interpreting the model's prediction using the CRP method is that we looked at conditioning on each layer, using the concept that was ranked as the most relevant for the prediction. In this report, we only show the most interesting features found for each dataset and model. We look at the same datasets and models as in the case of the LRP method. It was not trivial to apply CRP to ResNet since the residual connections caused the relevance to pass despite the masks being present. We were able to solve it by calculating the attributions twice, once conditioned on the concept we wanted, and once conditioned on a layer with no concepts, which allowed us to get exactly the relevance that was passed there through the residual connections.

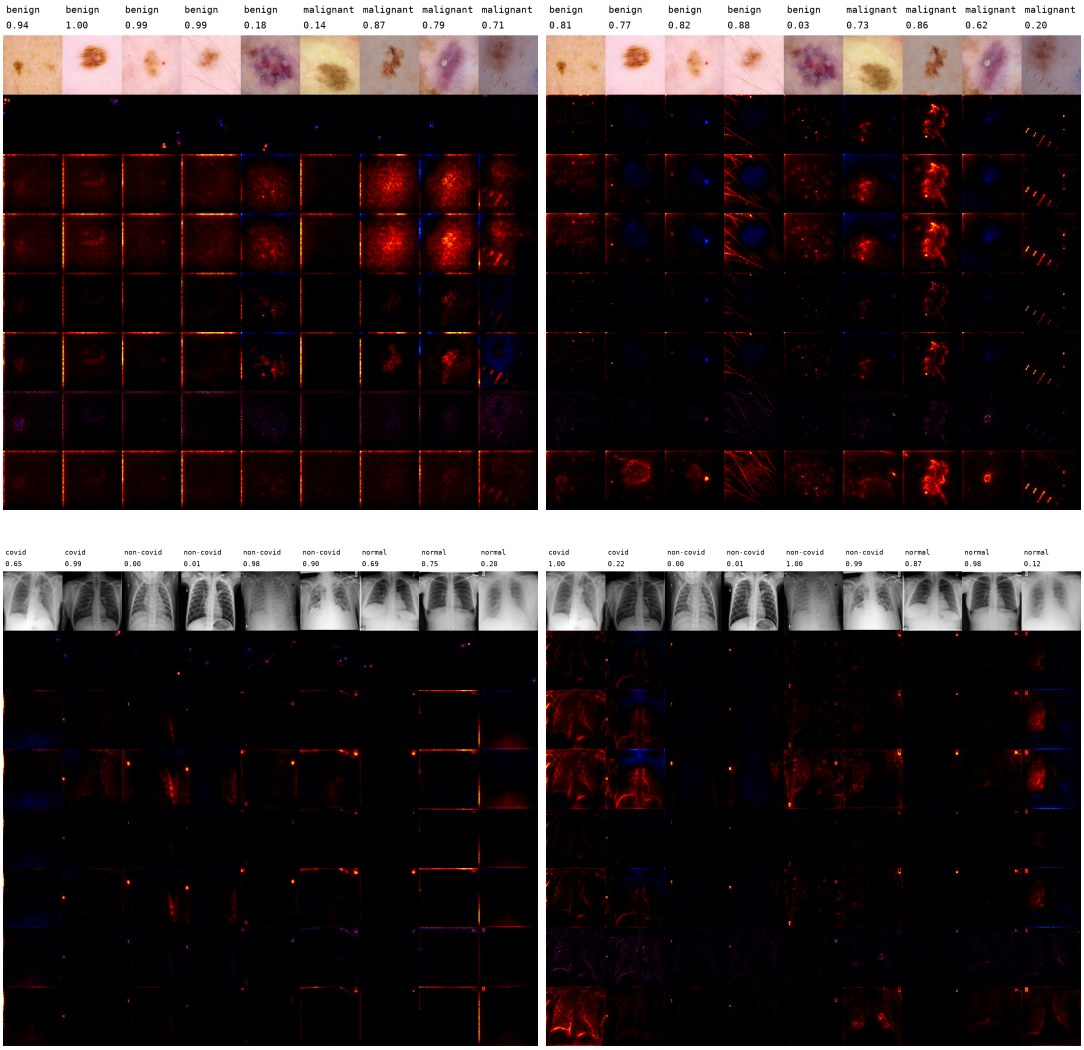

Figure 6: Melanoma LRP results (top left: ResNet50, top right: VGG16) and COVID LRP results (bottom left: ResNet50, bottom right: VGG16). On each image the rows are: class name, prediction on the class, input image, LRP explanations with composite: EpsilonGammaBox, EpsilonPlus, EpsilonPlusFlat, EpsilonAlpha2Beta1, EpsilonAlpha2Beta1Flat, GuidedBackprop, ExcitationBackprop.

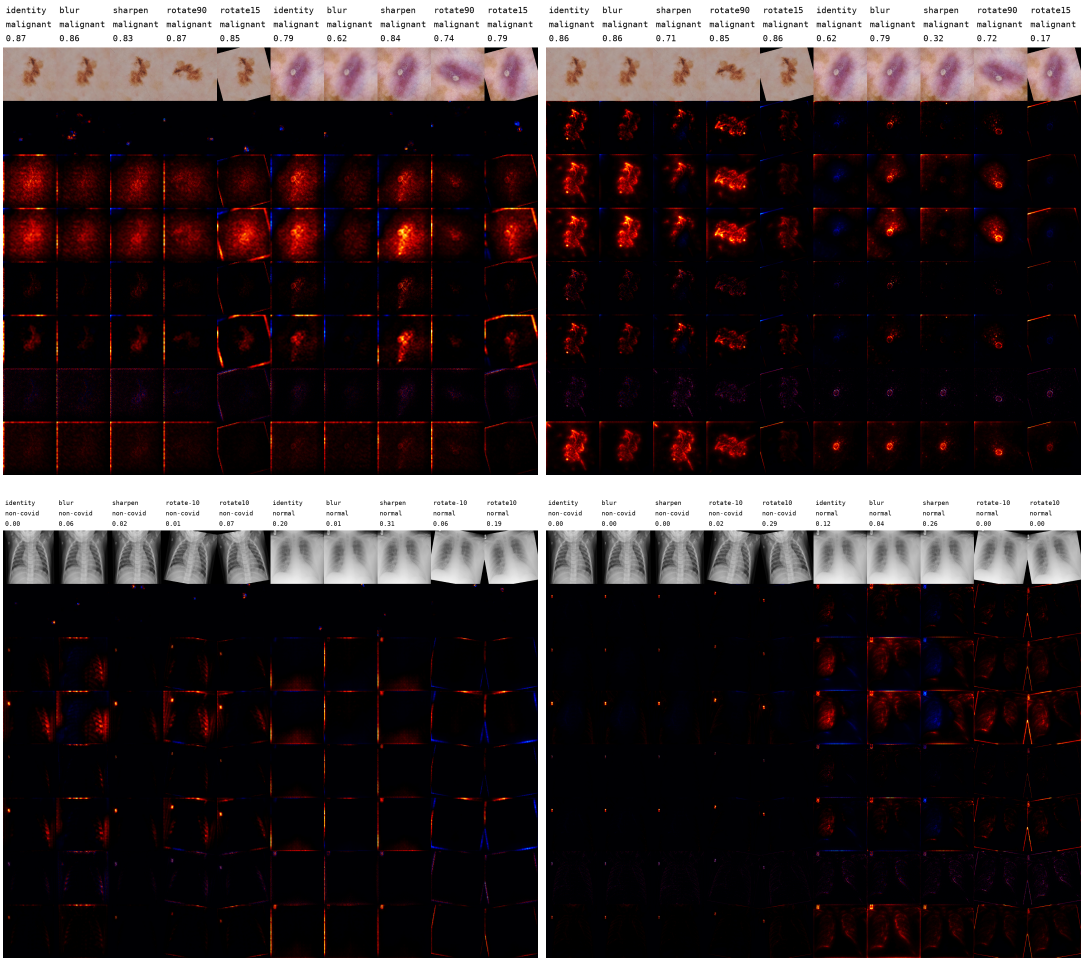

Figure 7: Results for LRP after input augmentation. Melanoma LRP results (top left: ResNet50, top right: VGG16) and COVID LRP results (bottom left: ResNet50, bottom right: VGG16). On each image the rows are: augmentation name, class name, prediction on the class, input image, LRP explainations with composite: EpsilonGammaBox, EpsilonPlus, EpsilonPlusFlat, EpsilonAlpha2Beta1, EpsilonAlpha2Beta1Flat, GuidedBackprop, ExcitationBackprop.