A European Cancer Image Platform Linked to Biological and Health Data for
Next-Generation Artificial Intelligence and Precision Medicine in Oncology

Deliverable D5.5:

Toolbox for interpretable AI in cancer imaging

Reference	D5.5_ EuCanImage_UM_28092023
Lead Beneficiary	UM
Author(s)	Zohaib Salahuddin, Henry Woodruff, Philippe Lambin
Dissemination level	Public
Type	Report
Official Delivery Date	28/09/2023
Date of validation of the WP leader	28/09/2023
Date of validation by the Project Coordinator	28/09/2023
Project Coordinator Signature	

EuCanImage is funded by the European Union’s H2020 Framework
Under Grant Agreement No 952103

Version log

Issue Date	Version	Involved	Comments
26/09/2023	v0.1	Zohaib Salahuddin	First Draft
26/09/2023	v0.2	Henry Woodruff, Philippe Lambin	Edits
26/09/2023	v0.3	Xènia Puig	Comments on the 1st draft
27/09/2023	v0.4	Zohaib Salahuddin	Changes
28/09/2023	final	Xènia Puig, Oliver Díaz, Karim Lekadir	Final and revised version

Executive Summary

Artificial intelligence has demonstrated promising performance on a variety of tasks for healthcare using medical images. However, there is a need to understand the decision-making process of the AI models for legal, ethical, and troubleshooting purposes. This document outlines the development of an interpretability toolbox aimed at improving the transparency of both handcrafted radiomics and deep learning solutions in medical imaging. For handcrafted radiomics, the toolbox offers tools for generating Shapley Additive Explanations, Local Interpretable Model-Agnostic Explanations, and tabular counterfactuals applicable to any machine learning model. For deep learning, the toolbox provides functionality to create attribution maps for both 2D and 3D classification models, along with an illustrative example of generating counterfactual image explanations. Additionally, the toolbox extends the counterfactual framework to incorporate layer-wise relevance propagation, allowing for the inclusion of clinical variables in decision-making. Finally, the document showcases an example platform built with Python and Streamlit, designed to validate these explanations within a clinical context, thereby evaluating their usability and value in the decision-making process. The interpretability toolbox is available at the following link: https://github.com/ZohaibS1995/radiomics_explainability_toolbox.

Table of Contents

Acronyms

Name	Abbreviation
Artificial Intelligence	AI
Deep Learning	DL
Machine Learning	ML
Handcrafted Radiomics	HCR
Shapley Additive Explanations	SHAP
Local Interpretable Model-agnostic Explanations	LIME
Layerwise Relevance Propagation	LRP
Shear wave Elastography	SWE

List of figures

1. Introduction

The increasing volume of medical imaging data poses challenges for radiologists and clinicians, leading to a growing need for tools to aid in diagnosis and decision-making. Deep learning (DL), a promising AI technology, has shown exceptional performance in medical imaging tasks but faces hurdles in adoption due to its "black-box" nature, which hinders transparency and compliance with regulations. Interpretability of DL systems is crucial, as it not only unravels the inner workings of algorithms and ensures compliance with legal requirements but also fosters clinical trust and reveals hidden insights in imaging data. Explainable artificial intelligence (XAI) seeks to make AI systems understandable to end-users [1]. Interpretability refers to techniques that elucidate why a DL model makes specific predictions in medical image analysis.

Handcrafted radiomics relies on manually designed features extracted from medical imaging data, involving expert-defined characteristics to describe regions of interest. In contrast, deep learning employs neural networks to automatically learn and extract intricate patterns and features directly from the raw image data, enabling more complex and data-driven representations. Both approaches have various techniques for interpretability, such as Shapley Additive Explanations [2] for handcrafted radiomics and methods like Grad-CAM [3] for deep learning, facilitating the understanding of model decisions and enhancing their clinical applicability.

In this document, we describe the interpretability toolbox that we have developed to enhance the explainability of both handcrafted radiomics and deep learning solutions. For handcrafted radiomics, we provide an interface to generate Shapley Additive Explanations, Local Interpretable Model-Agnostic Explanations, and tabular counterfactuals for any machine learning model. For deep learning, the toolbox offers functionality to generate attribution maps for 2D and 3D classification models. It also provides an example demonstrating how to generate counterfactual image explanations. The counterfactual framework is further extended to incorporate layer-wise relevance propagation [4] to include clinical variables in the decision-making process. Finally, the toolbox includes an example of a platform based on Python and Streamlit, designed to validate explanations in a clinical setting, thereby assessing the usability and added value of the explanations in the decision-making process.

2. Interpretability Methods

In this section, we briefly describe the different types of interpretability methods included in the toolbox as shown in figure 1. The toolbox supports interpretability methods for both handcrafted radiomics and deep learning.

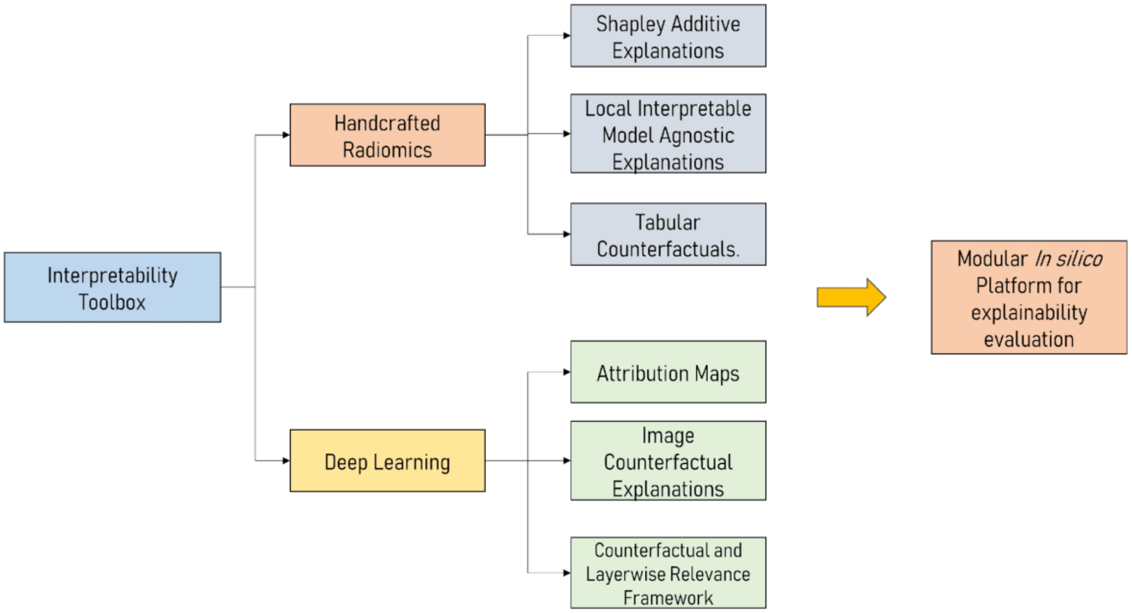

Interpretability Toolbox
Handcrafted Radiomics
Shapley Additive Explanations
Local Interpretable Model Agnostic Explanations
Tabular Counterfactuals.
Deep Learning
Attribution Maps
Image Counterfactual Explanations
Counterfactual and Layerwise Relevance Framework
Modular *In silico* Platform for explainability evaluation

Figure 1: The interpretability toolbox provides functionalities for generating explanations for both handcrafted radiomics and deep learning solutions along with a platform for quantitative evaluation of explanations in a clinical setting.

1. Shapley Additive Explanations

Shapley Additive Explanations (SHAP) is a method for understanding how individual features contribute to the predictions of a machine learning model [3]. SHAP is based on game theory, and it provides a way to fairly distribute the credit for a prediction to each feature. In the context of AI interpretability, SHAP can be used to answer the question: What is the impact of each feature on a particular model's prediction?

SHAP can be used to generate both local and global explanations, each of which has its own benefits. Local explanations explain the predictions of a specific instance or data point. SHAP can be used to understand why a particular model made a specific prediction for an individual observation. This is useful for understanding how the model works on a case-by-case basis. Global explanations provide an overview of model behavior across an entire dataset. SHAP can be used to generate summary statistics and feature importance rankings, which offer a holistic view of how each feature contributes to the model's predictions. Additionally, SHAP includes dependence plots, which visualize the relationship between a feature's value and its Shapley values.

2. Local Interpretable Model Agnostic Explanations

Local Interpretable Model-Agnostic Explanations (LIME) is a method for understanding why a machine learning model makes a particular prediction for a specific data point [5]. It does this by creating a simple, interpretable model that mimics the behavior of the original model at that point. LIME can be used to explain any type of machine learning model, regardless of its complexity.

3. Tabular Counterfactual Explanations

Diverse Counterfactual Explanations with Minimum Latent Effects (DICE-ML) is a method for generating counterfactual explanations for machine learning models [6]. Counterfactual explanations are examples of how a model's prediction would change if one or more of the input features were changed. Dice-ML aims to generate counterfactual explanations that are both diverse and meaningful. Dice-ML does this by perturbing the input features in a way that minimizes the extent of the perturbations. This ensures that the counterfactual explanations are faithful to the original model, but also that they are diverse enough to provide a comprehensive understanding of how the model works.

4. Attribution Maps

Attribution maps are essential tools for interpreting the decisions made by deep neural networks, highlighting which parts of an input contribute most to a model's prediction. These maps highlight the important regions of the input image for prediction. Some examples of popular attribution methods include GradCAM, Integrated Gradients, Input x Gradient, Guided GradCAM, and Guided Backpropagation. Attribution maps have a limitation such that they only highlight the region in the input image important for prediction but fail to show how these regions contribute towards the model’s output.

5. Counterfactual Explanations

Counterfactuals explanations are generated by applying minimal perturbation to the input image in such a manner that the prediction of the classifier was switched [1]. Counterfactual explanations go one step beyond the traditional heatmaps as they show by generating new samples that correspond to the change in prediction of the classifier.

We trained an MLP using the latent space representations of Variational Autoencoder as an input. We can exploit the continuity in the latent space of Variational Autoencoder to perturb the latent space and generate new images using the decoder. We followed the methodology in [7] to apply the modification to the semantically important pixels for classification using the gradient of the classifier.

6. Counterfactual and Layerwise Relevance Propagation Framework

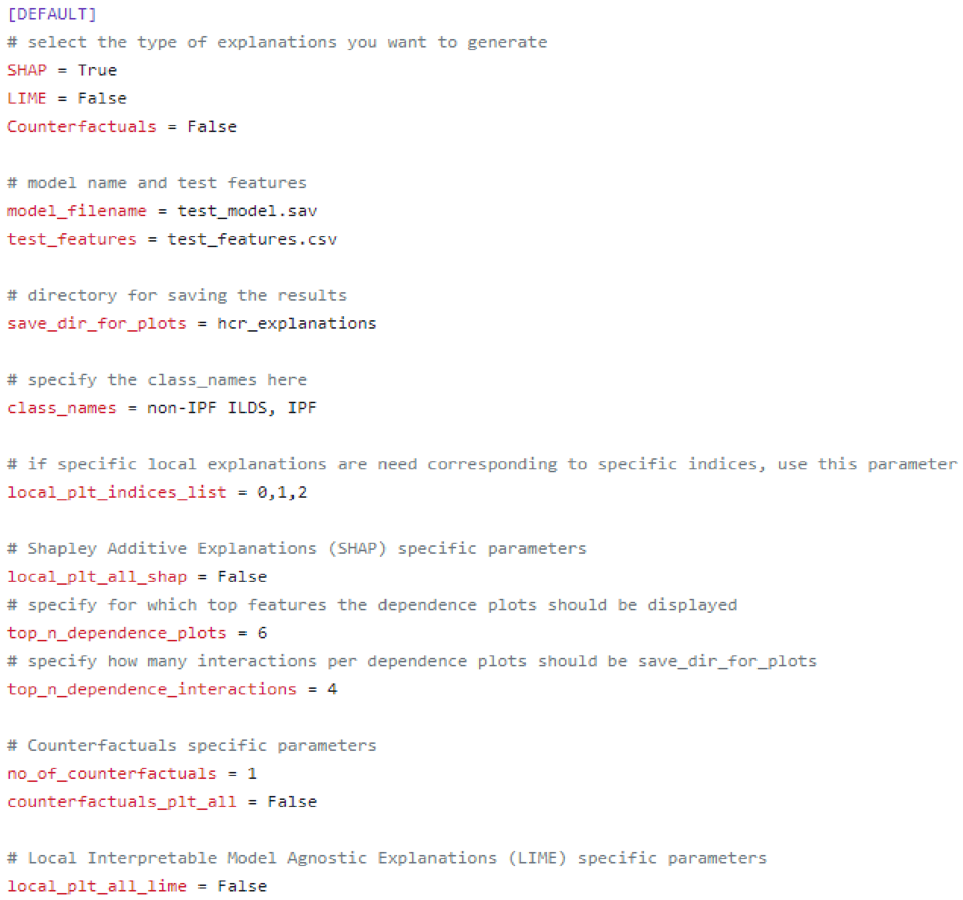
Layerwise relevance propagation (LRP) is an attribution method that calculates the contribution of each neuron by propagating the prediction backward based on relevance scores [8]. The total relevance at each layer of the neural network remains constant, starting from the last layer of the classifier.

Clinical variables can also be integrated within the deep learning network to build models based on both images and clinical variables. It is crucial to determine the contribution of each clinical variable and the images to the model’s prediction. Using LRP, we can assess the contribution of the image input and each clinical variable. Furthermore, image counterfactuals can be employed within the LRP framework to understand how changes in the input image affect the model’s prediction.

3. Interpretability Toolbox

1. Handcrafted Radiomics

The explanations generated for any type of handcrafted machine learning model are controlled by a *config.ini* file. The configuration file contains some default parameters. The path of the machine learning model saved in ".sav" format needs to be provided, along with the path of the test features file in ".csv" format that contains the features used to build the machine learning model. The last column of the test features should contain a column labeled "target" containing the labels. Figure 2 shows the different options that are present in the configuration file.



```
[DEFAULT]
# select the type of explanations you want to generate
SHAP = True
LIME = False
Counterfactuals = False

# model name and test features
model_filename = test_model.sav
test_features = test_features.csv

# directory for saving the results
save_dir_for_plots = hcr_explanations

# specify the class_names here
class_names = non-IPF ILDS, IPF

# if specific local explanations are need corresponding to specific indices, use this parameter
local_plt_indices_list = 0,1,2

# Shapley Additive Explanations (SHAP) specific parameters
local_plt_all_shap = False
# specify for which top features the dependence plots should be displayed
top_n_dependence_plots = 6
# specify how many interactions per dependence plots should be save_dir_for_plots
top_n_dependence_interactions = 4

# Counterfactuals specific parameters
no_of_counterfactuals = 1
counterfactuals_plt_all = False

# Local Interpretable Model Agnostic Explanations (LIME) specific parameters
local_plt_all_lime = False
```

Figure 2: The configuration file containing default input options for generating different types of explanations for the handcrafted radiomics classifier.

Figures 3 (A) and (B) show the SHAP global summary plots and SHAP dependence plots for a handcrafted radiomics model [9] that are generated using the interpretability toolbox. Figure 4 (A) and (B) show examples of local explanations using SHAP and LIME methods respectively that are generated using the interpretability toolbox.

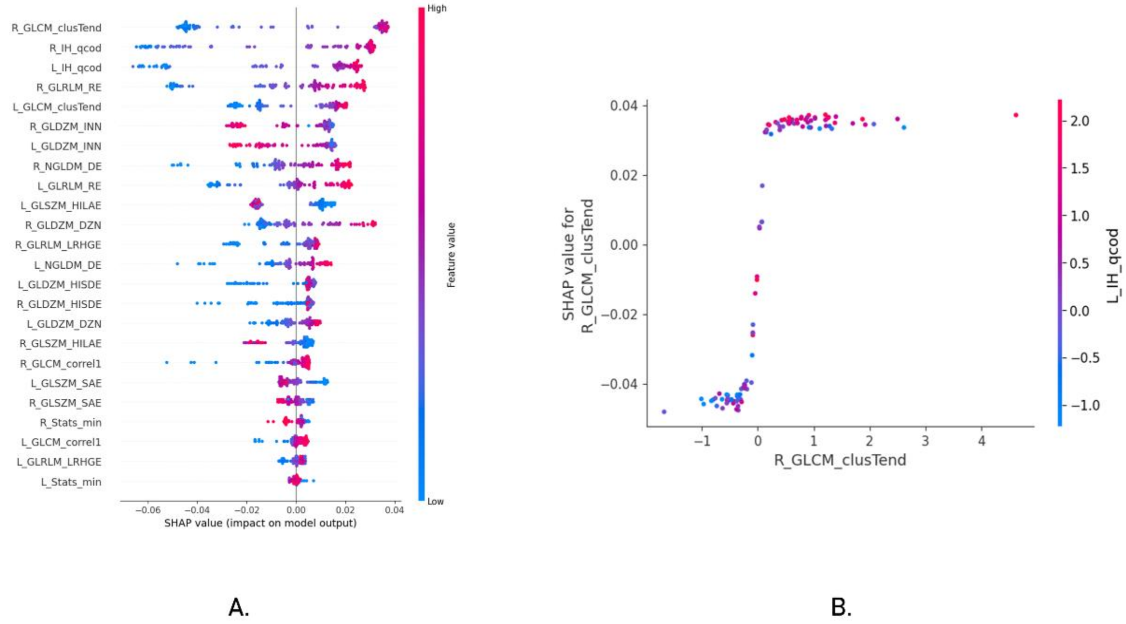


Figure 3: A) Global SHAP summary plot for a handcrafted radiomics model, B) Dependence plot for a handcrafted radiomics model.

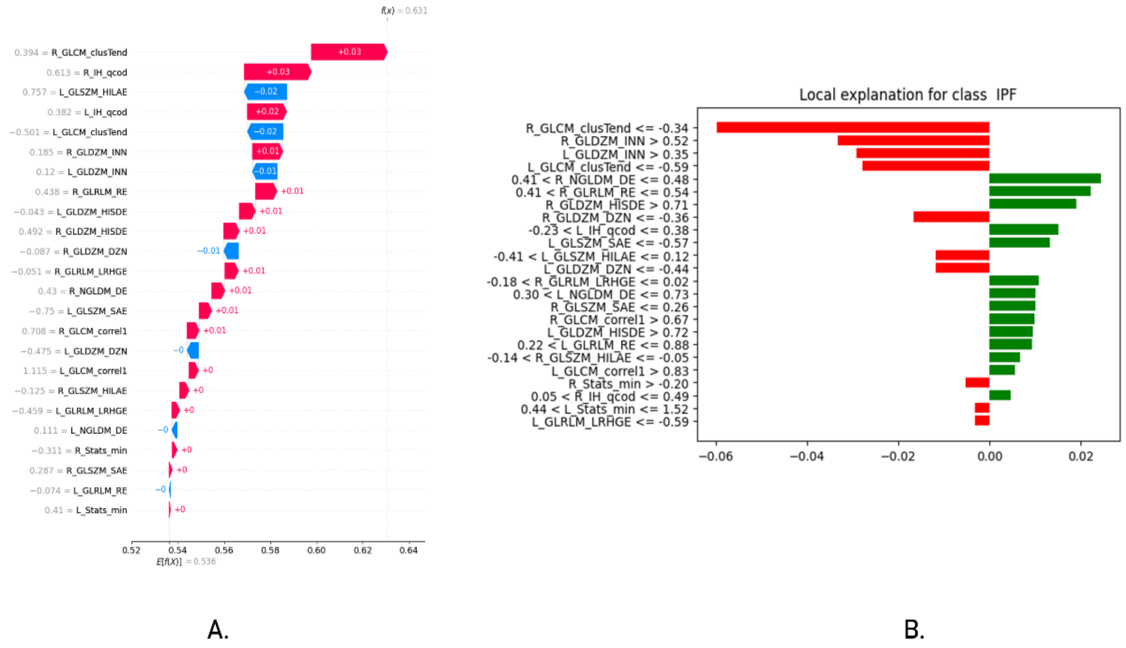


Figure 4: A) Local SHAP plot for a specific instance, B) Local LIME plot for a specific instance.

2. Deep Learning

1. Attribution Maps

The interpretability toolbox supports the following methods for generating attribution maps for PyTorch-based classifiers: GRADCAM, Guided-GRADCAM, Input x Gradient, Guided Backpropagation, and Integrated Gradients. These methods are configured using a config.ini file. The model definition and pre-processing function need to be edited in the main.py file. Figure 5 (A) and (B) show how the interpretability toolbox can be used to generate explanations for 2D medical images, and Figure 5 (C) shows an example of attribution maps for a 3D image.

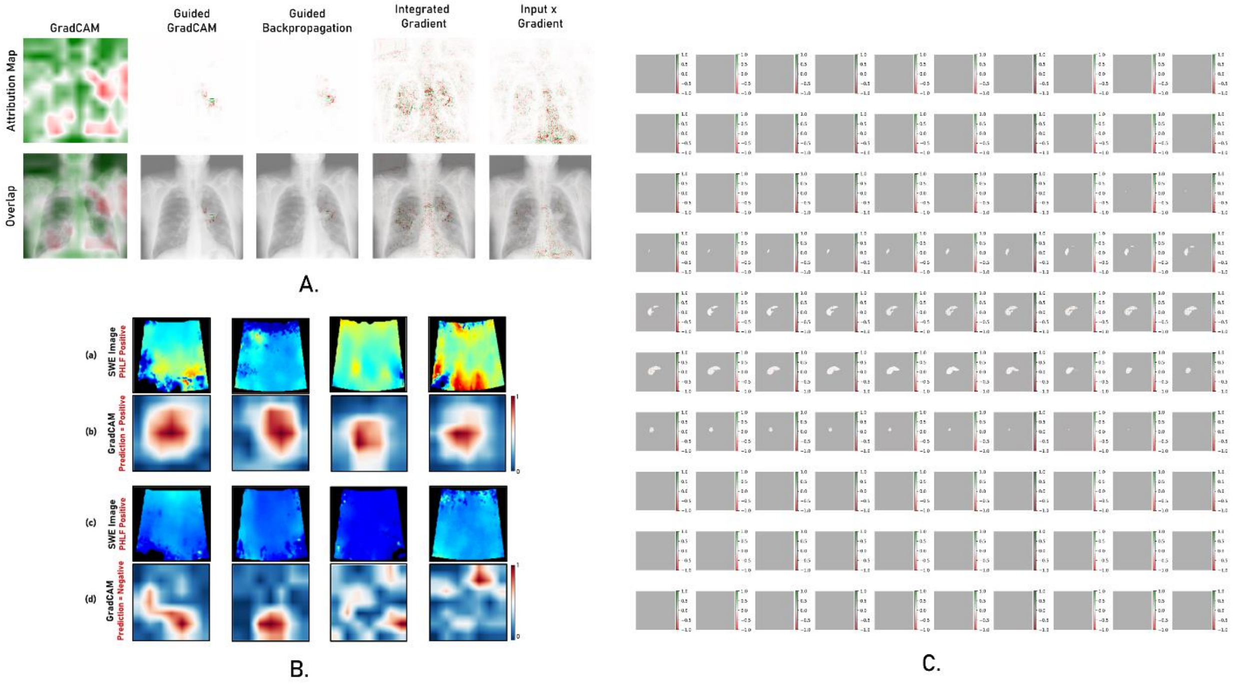


Figure 5: A) and B) show attribution maps for 2-dimensional X-ray and SWE images and C) shows attribution maps for a 3-dimensional CT image.

2. Counterfactual Explanation

We demonstrate the process of generating counterfactual explanations using a variational autoencoder for post-hepatectomy liver failure using 2-dimensional SWE images. The process for generating counterfactuals using the interpretability toolbox involves three steps. The first step is to train the variational autoencoder in an unsupervised manner to generate latent representations. The second step is to train the deep learning classifier for the task at hand. The third step is to generate the counterfactuals by traversing the latent space of the variational autoencoder with respect to the classifier. Four examples of counterfactual explanations for images with different predicted probabilities are shown in Fig. 6.

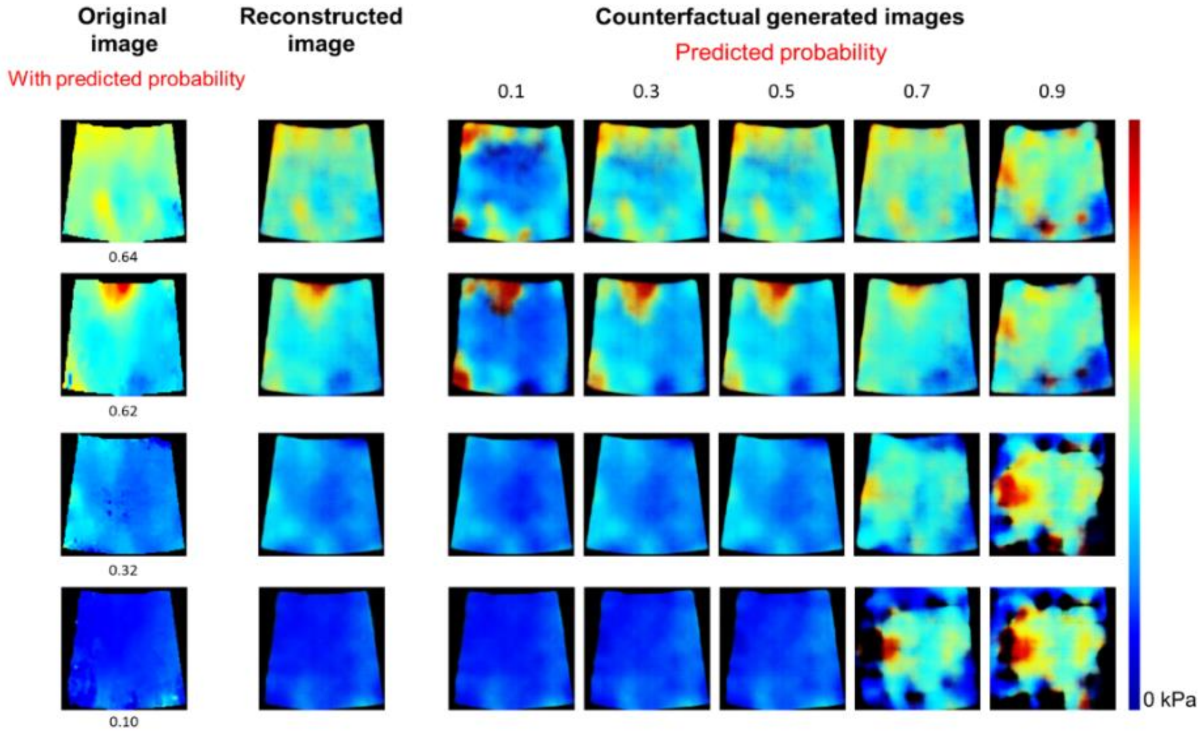
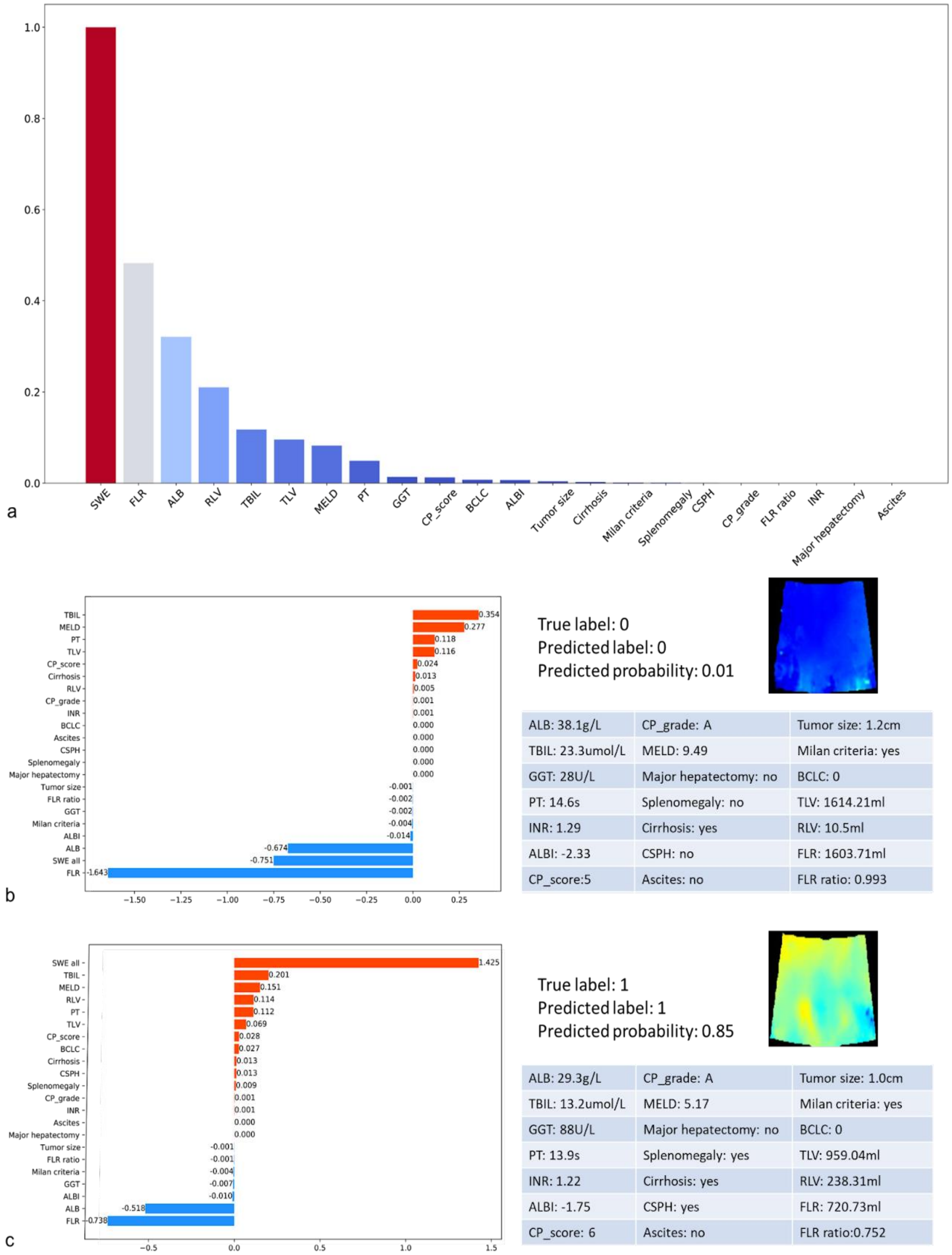


Figure 6: Four examples of counterfactual explanations generation for the VAE-MLP model. The first column showed the original images with the classifier's predicted probability shown below each image. The second column showed the corresponding reconstructed image. For the input image in the first column, our model generates a series of counterfactual images as explanations with predicted probabilities of 0.1, 0.3, 0.5, 0.7, and 0.9, which were shown in the right part of the figure.

3. Counterfactual and Layerwise Relevance Propagation Framework.

We demonstrate how clinical variables can be incorporated into the deep learning workflow in the interpretability toolbox. The deep learning model that makes its predictions based on images and clinical variables can be explained using the counterfactual and layerwise relevance propagation framework. The contribution of each image and clinical variable, both in the local and global sense, is determined using the layerwise relevance propagation framework. Counterfactual explanations are generated by traversing the latent space of the VAE while holding the clinical variables constant to determine how changes in the image impact the model's prediction while keeping the clinical variables constant.

A Global LRP analysis of a model identified SWE, future liver remnant volume (FLR), albumin (ALB) as the most important features for symptomatic PHLF prediction. Other variables such as resected liver volume (LRV), total bilirubin (TBIL), total liver volume (TLV), model for end-stage liver disease (MELD) score, prothrombin time (PT), also contributed to the prediction (Fig. 7a). Figs. 7b and 7c show LRP local bar plots for two test cases. Fig. 7b shows a case without symptomatic PHLF that had been classified by the model. The plot shows that FLR, SWE and ALBI contributed most to the negative prediction. Fig. 7c shows a case with symptomatic PHLF that has been classified correctly by the model. The plot shows that SWE contributed most to the positive prediction and FLR and ALB contributed most to the negative prediction.



True label: 0
Predicted label: 0
Predicted probability: 0.01

ALB: 38.1g/L	CP_grade: A	Tumor size: 1.2cm
TBIL: 23.3umol/L	MELD: 9.49	Milan criteria: yes
GGT: 28U/L	Major hepatectomy: no	BCLC: 0
PT: 14.6s	Splenomegaly: no	TLV: 1614.21ml
INR: 1.29	Cirrhosis: yes	RLV: 10.5ml
ALBI: -2.33	CSPH: no	FLR: 1603.71ml
CP_score:5	Ascites: no	FLR ratio: 0.993

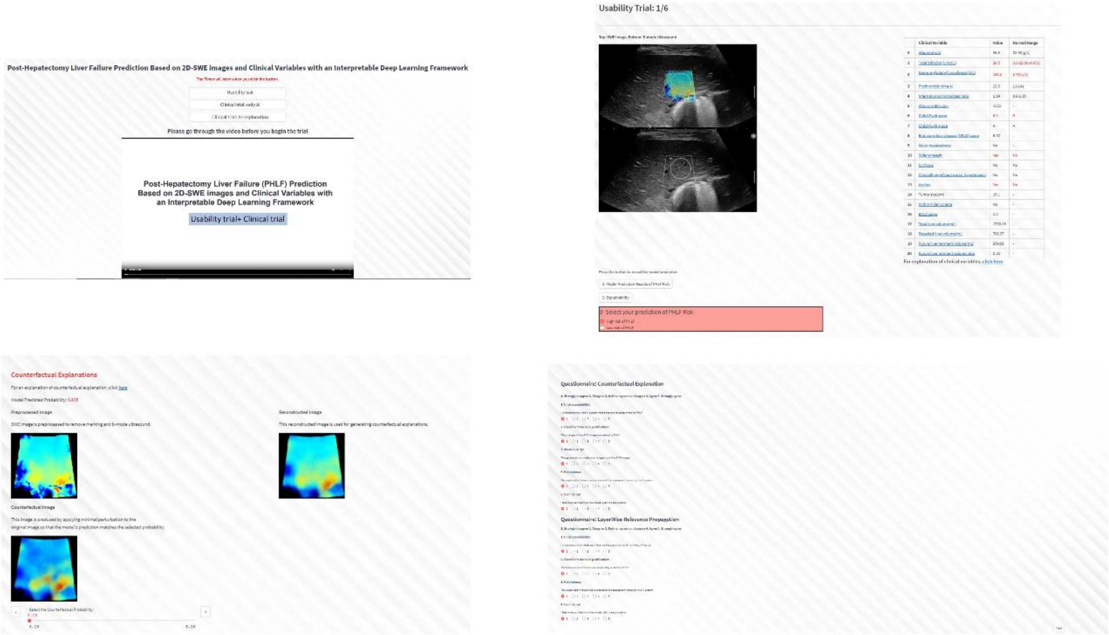
True label: 1
Predicted label: 1
Predicted probability: 0.85

ALB: 29.3g/L	CP_grade: A	Tumor size: 1.0cm
TBIL: 13.2umol/L	MELD: 5.17	Milan criteria: yes
GGT: 88U/L	Major hepatectomy: no	BCLC: 0
PT: 13.9s	Splenomegaly: yes	TLV: 959.04ml
INR: 1.22	Cirrhosis: yes	RLV: 238.31ml
ALBI: -1.75	CSPH: yes	FLR: 720.73ml
CP_score: 6	Ascites: no	FLR ratio:0.752

Figure 7: (a) The global layer-wise relevance propagation showing different feature contribution. (b)The local layer-wise relevance bar plot showing the feature contributions for a case without symptomatic PHLF. (c) The local layer-wise relevance bar plot showing the feature contributions for a case with symptomatic PHLF. ALB, albumin;TBIL, total bilirubin; GGT, gamma-glutamyl transferase; PT, prothrombin time; INR, international normalized ratio; ALBI: Albumin-Bilirubin; CP_score: Child-Pugh score; CP_grade: Child-Pugh grade; MELD: model for end-stage liver disease; CSPH: Clinically significant portal hypertension; BCLC, Barcelona Clinic Liver Cancer; TLV: total liver volume; RLV: resected liver volume; FLR: future liver remnant volume; PHLF: post-hepatectomy liver failure.

4. Platform for Validation of Explanations

It is important to carry out a usability evaluation of explanations as well as an evaluation of explanations in a clinical workflow. The interpretability toolbox contains an example platform for the clinical validation and usability assessment of various explanation types. The platform has been developed in a modular fashion with a config file. The config path contains paths to different types of explanations. The following link hosts an example of the platform for the validation of the counterfactual and layerwise relevance propagation framework: https://st-trial-hbhdzvtdqlr.streamlit.app/. Figure 8 shows different pages from the modular platform.



Post-Hepatectomy Liver Failure Prediction Based on 2D-SWE images and Clinical Variables with an Interpretable Deep Learning Framework

Please go through the video before you begin the trial

Post-Hepatectomy Liver Failure (PHLF) Prediction Based on 2D-SWE images and Clinical Variables with an Interpretable Deep Learning Framework

Usability trial+ Clinical trial

Usability Trial: 1/6

For explanation of clinical variables, click here

3- Select your prediction of PHLF Risk

Counterfactual Explanations

For an explanation of counterfactual explanation, click here

Model Predicted Probability: 0.808

Preprocessed Image

SWE image is preprocessed to remove marking and b-mode ultrasound.

Reconstructed Image

This reconstructed image is used for generating counterfactual explanations.

Counterfactual Image

This image is produced by applying minimal perturbation to the original image so that the model's prediction matches the selected probability.

Select the Counterfactual Probability:

Questionnaire: Counterfactual Explanation

Questionnaire: LayerWise Relevance Propagation

5. Next Steps

In this document, we present a comprehensive interpretability toolbox that can be used to generate explanations for both handcrafted radiomics and deep learning solutions. In the future, additional explainability tools will be incorporated into this toolbox. The handcrafted radiomics toolbox will be integrated into the virtual research environment to generate explanations for any handcrafted radiomics model.

6. References

[1] Salahuddin, Z., Woodruff, H.C., Chatterjee, A. and Lambin, P., 2022. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Computers in biology and medicine, 140, p.105111.

[2] Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[3] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

[4] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R. and Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, *10*(7), p.e0130140.

[5] Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. " 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* (pp. 1135-1144).

[6] Mothilal, R.K., Sharma, A. and Tan, C., 2020, January. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 607-617).

[7] Cohen, J.P., Brooks, R., En, S., Zucker, E., Pareek, A., Lungren, M.P. and Chaudhari, A., 2021, August. Gifsplanation via latent shift: a simple autoencoder approach to counterfactual generation for chest x-rays. In *Medical Imaging with Deep Learning* (pp. 74-104). PMLR.

[8] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R. and Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, *10*(7), p.e0130140.

[9] Salahuddin, Z., Frix, A.N., Yan, C., Wu, G., Woodruff, H.C., Gietema, H., Meunier, P., Louis, R., Guiot, J. and Lambin, P., 2022. Diagnosis of idiopathic pulmonary fibrosis in high-resolution computed tomography scans using a combination of handcrafted radiomics and deep learning. *Frontiers in medicine*, *9*, p.915243.