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

Deliverable D5.2:

Cancer radiomics extraction and selection pipeline

Reference	D5.2_ EuCanImage_UB_v1
Lead Beneficiary	UB
Author(s)	Kaisar Kushibar, Socayna Jouide
Dissemination level	Public
Type	Report
Official Delivery Date	31 March 2022
Date of validation of the WP leader	31 March 2022
Date of validation by the Project Coordinator	31 March 2022
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
02/03/2022	v0.1	Kaisar Kushibar	First Draft
10/03/2022	v0.2	Socayna Jouide	Added VRE example visuals
20/03/2022	v0.3	Karim Lekadir	Feedback
30/03/2022	v1	Karim Lekadir, Isabell Tributsch	Revised and corrected final version.

Executive Summary

Radiomics extraction is an essential part of Machine Learning pipelines applied to Cancer Imaging. This document presents the progress done on this front for an open-source and scalable radiomics extraction tool.

In this deliverable, detailed functionality and specifics of the radiomics extraction process are provided. Also, a general overview of the preliminary steps such as lesion localization and segmentation tools are described. Moreover, one use-case application that includes radiomics feature selection pipeline for Breast Cancer Treatment Response Prediction using Magnetic Resonance Images is demonstrated.

This task is the initial step towards answering the research question raised in Use-case 7 of the EuCanImage project: Could AI tools enable de-escalating neoadjuvant systemic therapy (NST) in patients highly likely to achieve a pathological complete response (pCR)? Both the radiomics extraction and treatment response prediction tools have been integrated and publicly available within the AI Virtual Research Environment (AI-VRE) of the EuCanImage platform at https://vre.eucanimage.eu.

Table of Contents

Acronyms

Name	Abbreviation
Artificial Intelligence	AI
Deep Learning	DL
Machine Learning	ML
Virtual Research Environment	VRE
Magnetic Resonance Imaging	MRI
Mammography	MG
Computed Tomography	CT
Ultrasound	US
Convolutional Neural Network	CNN
Region of Interest	ROI

1 Introduction

Early detection of cancer is a crucial step to increase survival rates and patient's quality of life. Screening programmes provide additional aid to identify asymptomatic or early-stage cancer subjects. Radiological non-invasive imaging modalities such as Mammography (MG), Magnetic Resonance Imaging (MRI), Computed Tomography (CT), and Ultrasound (US) imaging techniques are widely used for diagnosis, treatment planning, and disease monitoring. Artificial Intelligence (AI) has become one of the leading research fields in oncology with the aim to assist radiologists and medical doctors with decision support systems. Due to the large number of modalities and organs it is crucial to have a universal computational framework where a commonly used and standardised basis for any AI tool is already developed.

Radiomics extraction and selection is one of these common grounds which is shared across all ML approaches regardless of the differences in modalities and organs. However, due to the differences in imaging techniques such as dimensionality (2D, 3D, 4D) and resolution (mm, μm) the radiomics extraction pipelines are often rewritten. Considering the various modalities and organs addressed in EuCanImage such as breast, colorectal, and liver with a variety of imaging modalities, this project provides a tool for radiomics extraction and selection. EuCanImage strives for building a world-wide open-access platform (i.e., VRE) where this tool has been integrated and accessible to the public in a cloud computing environment.

In this document, we demonstrate the general pipeline that leads to radiomics extraction and selection on breast cancer MG images based on Use-Case 8 of the EuCanImage project. Then, the application of the tool is assessed using an example of Use-Case 7 by building an ML model that predicts breast cancer neoadjuvant treatment response prediction using pre-treatment and single time point MRI images.

1.1 General Pipeline: Radiomics analysis in Cancer Imaging (Use-Case 8: Cancer diagnosis)

After image acquisition, there are some common prerequisites that need to be addressed before performing the radiomics and ML analysis on given images. Figure 1 illustrates the overall pipeline for radiomics analysis and machine learning for breast cancer diagnosis on screening and clinical mammogram images.

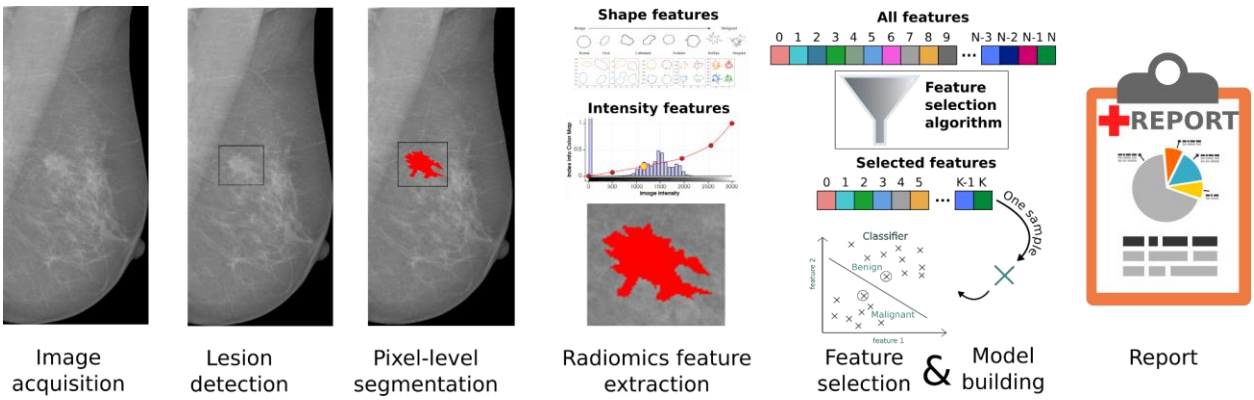

Figure 1: Illustration of full pipeline for ML using Radiomics for breast cancer detection

In this example for breast mass classification, two preliminary stages are followed after image acquisition: lesion detection and pixel-level segmentation of the detected lesion. Both types of annotations are often done manually by an expert radiologist, which is a time consuming and laborious task. Hence, automating these preliminary steps is critical in developing decision support systems. Accordingly, these two tools have been developed based on deep-learning models using multi-centre, multi-vendor public datasets. Table 1 shows acquired public datasets used in building both of the annotation tools.

Table 1: Publicly available datasets used in mass detection and segmentation AI models' development.

Dataset	OPTIMAM Hologic	OPTIMAM Siemens	OPTIMAM GE	OPTIMAM Philips	INbreast Siemens	BCDR
Cases	1924	65	45	208	50	334
Images	3446	120	83	407	107	886
Annotation type	Bounding box	Bounding box	Bounding box	Bounding box	Segmentation	Segmentation

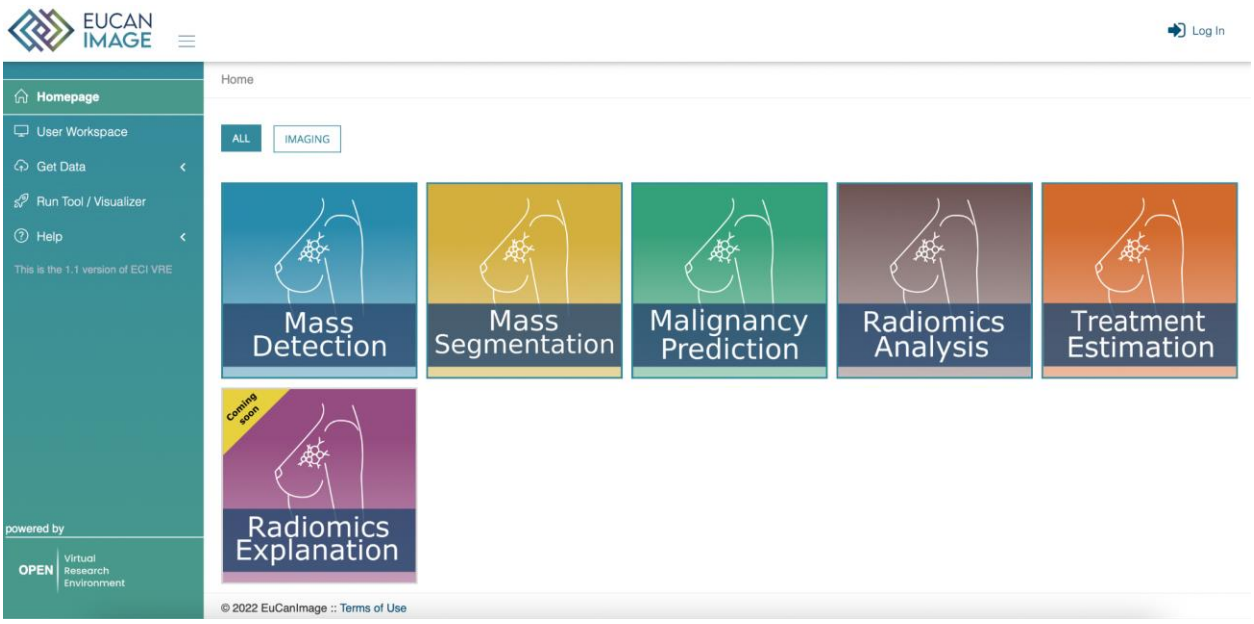

Figure 2: Homepage of the VRE. In this page, the available radiomics analysis tools are listed and can be assessed. Other available or under development tools can also be viewed with the coming soon tag.

Automatic annotation tools for breast mass detection and segmentation are integrated into the VRE platform of EuCanImage. Figure 2 provides an overview of the current state of the VRE homepage and the tools included. Each tool can be accessed by (i) the homepage (Figure 2), (ii) the "User workspace" tab (Figure 3) if the appropriate file types are chosen, i.e. the file types that each tool accepts as input, and (iii) the "Run Tool/Visualizer" tab (https://vre.eucanimage.eu/vre/launch/) (Figure 4). Each tool has its own front page (Figure 5), where it can be configured and when the user provides the adequate input and presses

the Compute button, the page will be redirected to the "User Workspace". The status of the job changes from pending to running to finishing. When the job finishes, the result(s) appears under the given experiment name and can be downloaded from there (Figure 6). The next section dives into the internal structures of each tool and the model training procedures that are used to achieve robust automated annotation tools.

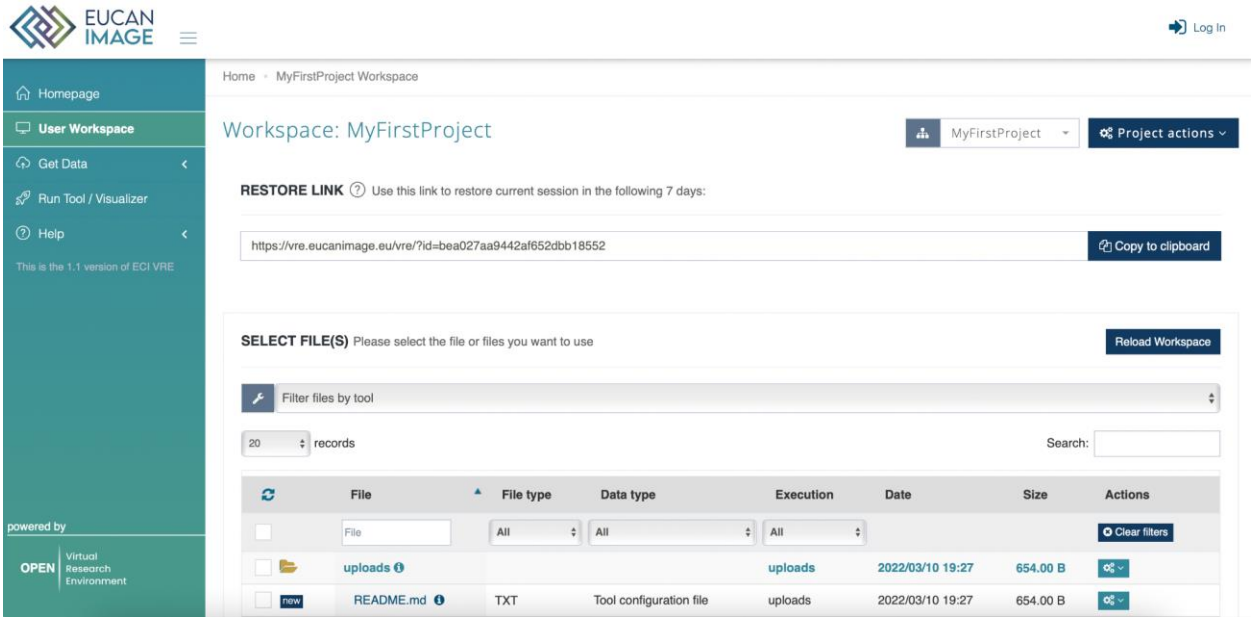

Figure 3: "User workspace" tab. In this page the user can find the previously uploaded data and the results of the experiments. In addition, these files can be selected to see available tools to process them. Lastly, the running jobs can be monitored while they change status from pending to running to finishing.

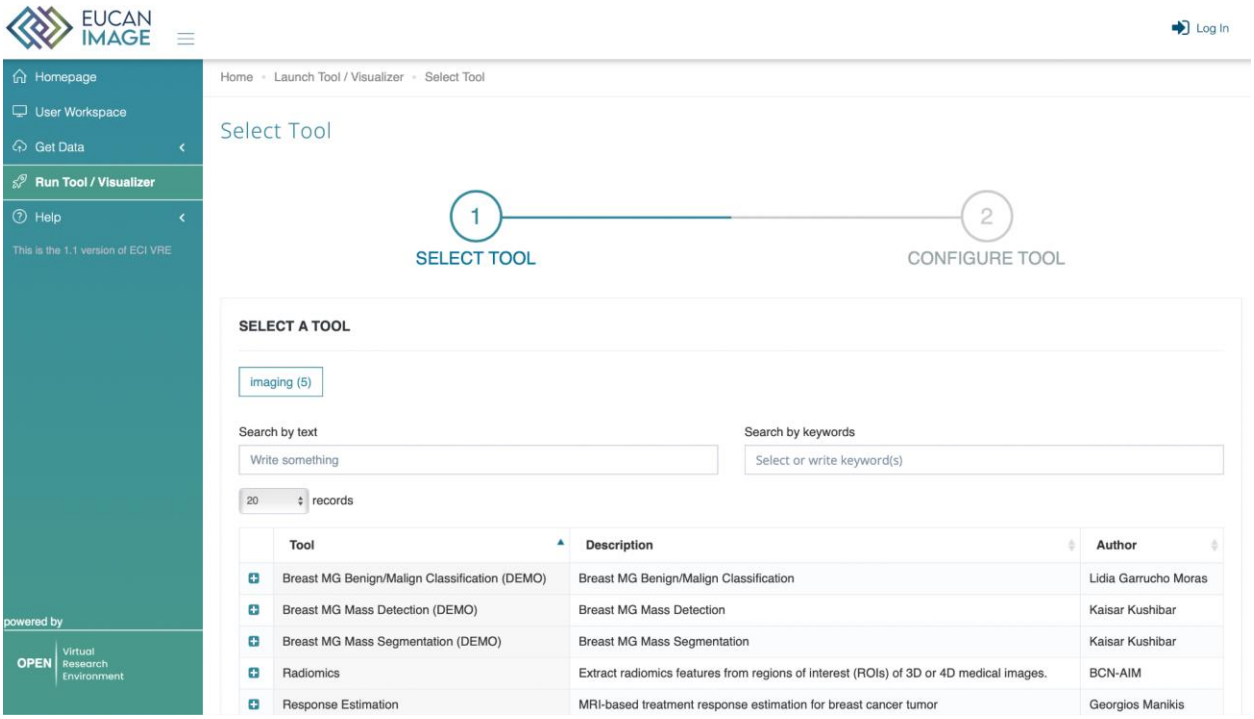

Figure 4: The Run tool/visualizer tab of the VRE. In this page, the available tools are listed and can be assessed.

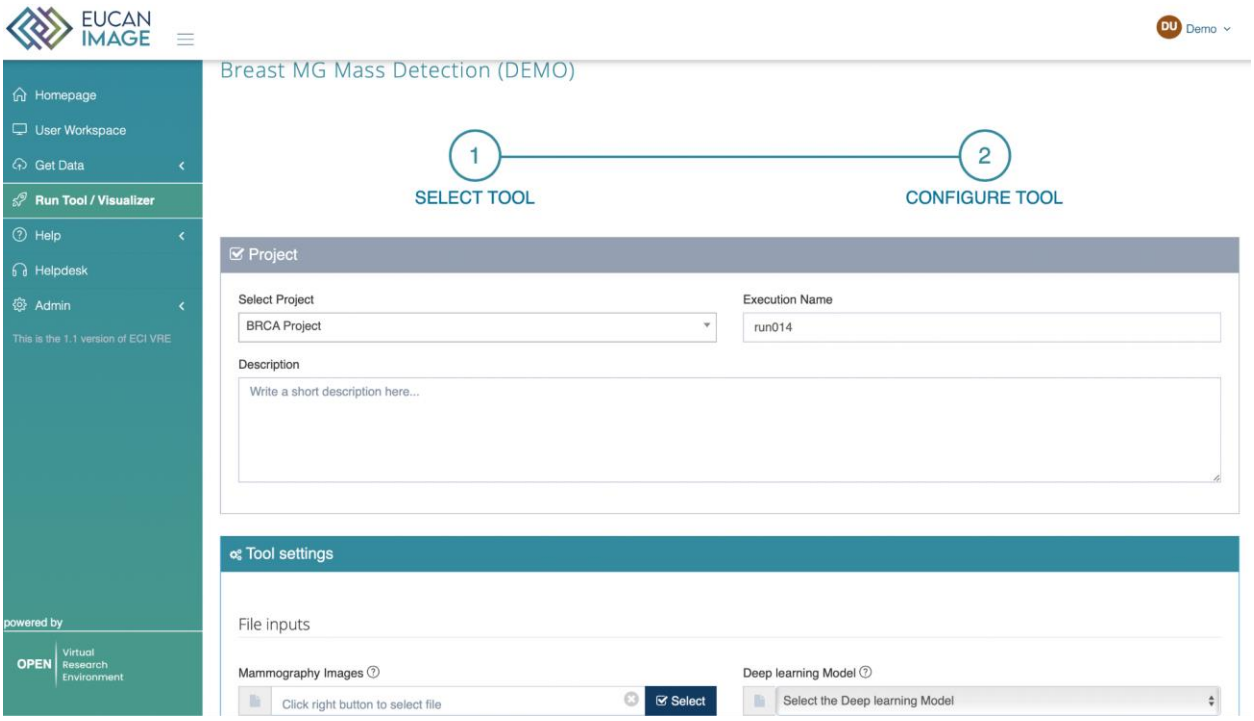

Figure 5: Main page and configuration settings for the breast mass detection tool. The input dataset must be selected and the option for the deep learning model must be defined prior to the tool's execution.

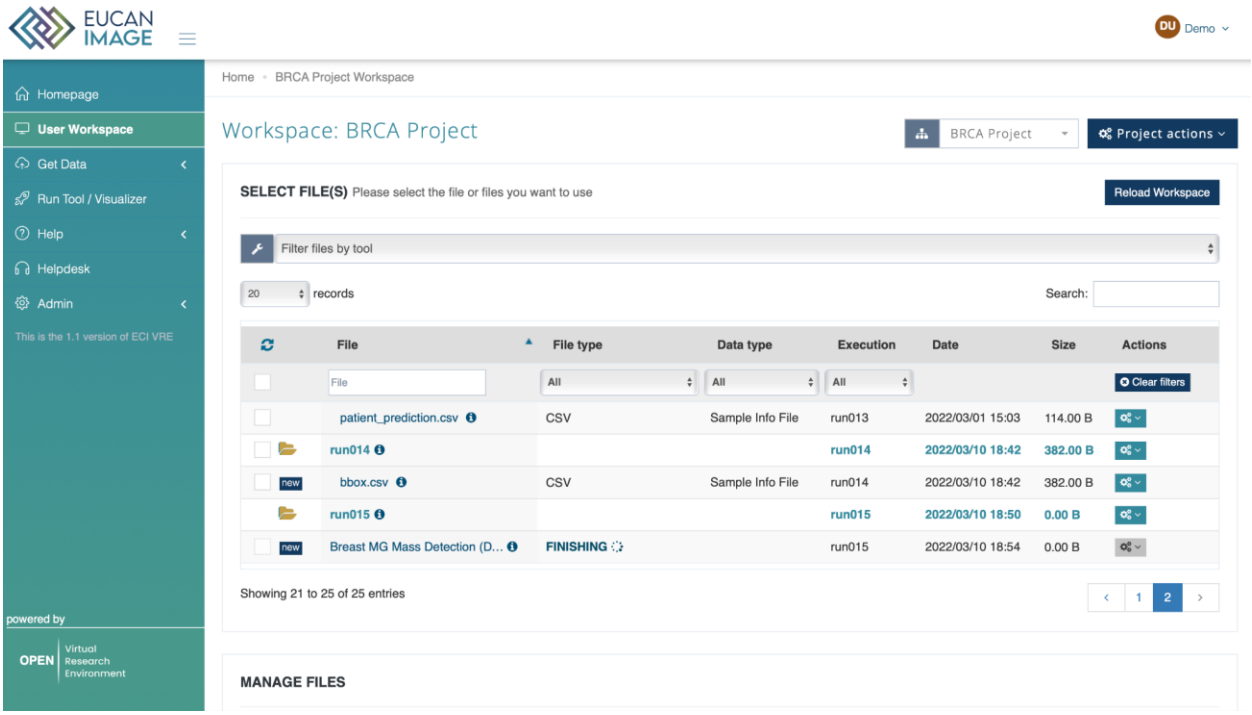

Figure 6: "User workspace" tab. After running jobs, this section of the VRE displays the progress of it and allows a monitorization of the execution and the possibility of downloading the results.

1.2 AI based tools to automate prior processes to radiomics extraction

Breast mass detection

The initial step after image acquisition is spatial localisation of breast masses in high-resolution MG images. EuCanImage data platform will be populated with data coming from different clinical centres that employ different scanner manufacturers. When building a robust AI solution, two important factors should be taken into account that are the predominant failure points: 1) domain shift – differences in scanner manufacturers and image acquisition protocols; and 2) dataset shift – differences in cohort specific distributions, e.g. lesion size, lesion aspect ratio. Both of these types of obstacles for AI should be taken into account at an early stage. Therefore, **domain generalisation** is a crucial step in building an AI solution.

The proposed mass detection AI tool addresses these major issues by building a robust deep learning based model. The following model training pipeline has been proposed (see Figure 7) [1]. First, scale intensity standardisation, to bring the histogram distributions of MG images into a standardised template using the algorithm proposed in [2]. Second, during training, along with standard data augmentation techniques such as flipping and rotation, Cutout [3] and RandConv [4] are applied to increase domain variability in the training images. Particularly, RandConv method allows transforming the texture of MG images while preserving the anatomical structures. Next, MixStyle [5] layers are added within ResNet50 CNN architecture that further increases domain variability in the deep feature level. Moreover, modern Transformer based architecture is employed, namely Deformable DETR [6], that utilises self-attention mechanisms that rely less on inductive biases that are dominant in traditional CNN based architectures. Then, the final model becomes less dependent on the

intensity differences caused by a variety of acquisition protocols and scanners. This approach is tested on multi-centre data and achieves state-of-the-art results.

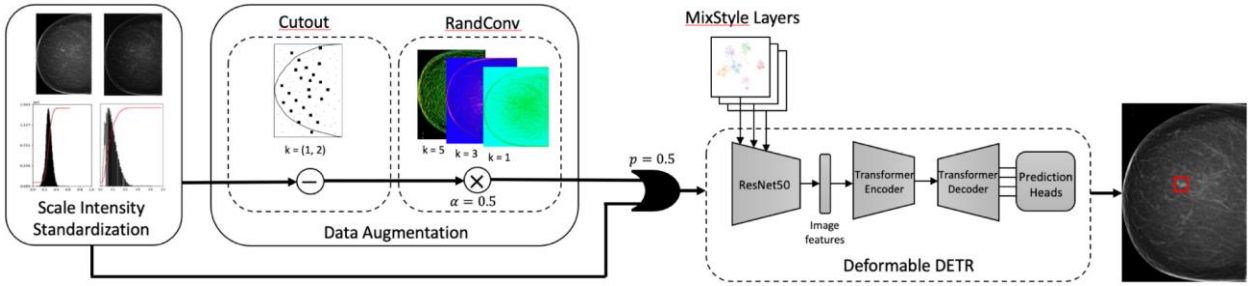

Figure 7: Scanner and domain shift robust breast mass detection pipeline. Figure courtesy of [1].

Breast mass segmentation

Once the masses are localised, accurate segmentation masks are generated using a deep learning based model that is shown in Figure 8. Pixel-accurate segmentation is crucial to extract radiomics features that are not affected by healthy tissue and the true shape of a lesion is well outlined. This model is also trained using multi-centre data coming from INbreast and BCDR public datasets (Table 1) that have fine segmentation masks. The larger dataset OPTIMAM was not included due to the absence of lesion segmentations. Unlike the detection model, the number of training images is much less, hence, the network is shallower. However, to make this solution work, multiple additional techniques are utilised to improve the capacity of the network. First, the well-known U-Net architecture is used that has skip-connections between encoder and decoder blocks that are beneficial to avoid common problems such as vanishing gradient as well as benefiting from multi-level feature fusion. Furthermore, encoder blocks are based on residual connections and decoder blocks are equipped with attention mechanisms. Then, one of the most important performance gains is achieved using deep supervision strategy, where segmentation heads are attached after each encoder and decoder blocks. This allows the network to be supervised in every level of the network. This model also achieves competitive results with the state-of-the-art for mass segmentation.

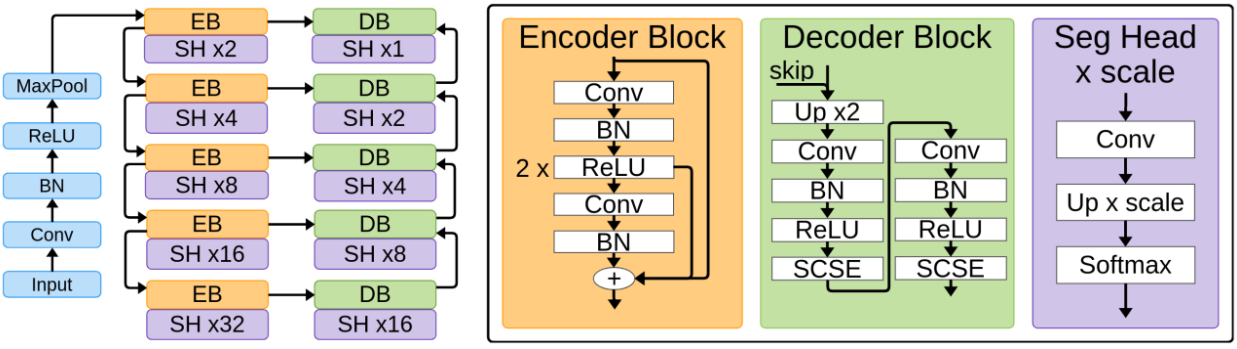

Figure 8: Breast mass segmentation network that is built on top of U-Net like architecture. Encoder Block (EB) in orange and Decoder Block (DB) in green have internal structures as depicted in the boxes below with corresponding colours. Ten Segmentation Heads (SH) are attached after each layer output with an up-scaling factor depending on the depth of the layer. SCSE - Squeeze and Excitation attention module. BN - Batch Normalisation.

2 Radiomics analysis tool

Radiomics analysis has shown great potential in both MG and MRI modalities for diagnosis, treatment response, survival prediction, and molecular-subtype classification in breast cancer [7,8]. The radiomics analysis tool performs the extraction of a high number of quantitative features to characterise shape, intensity and texture of the Region of Interest (ROI), known as **radiomics**. The ROIs are delineated manually or using fully or semi-automatic tools. As it has been shown before, lesion localisation and segmentation tools are already integrated within the VRE of EuCanImage for MG modality. Currently, for the MRI modality, the radiomics analysis tool expects already delineated lesion ROIs. However, as part of the Use-Case 6, development of automatic detection and segmentation tools is scheduled and will be initiated upon availability of training data from the clinical partners involved in the project.

2.1 Radiomics features

The radiomics analysis tool of the EuCanImage platform is based on the open-source python-based PyRadiomics platform version 3.0.1 and is developed in Python. In total, 105 features per ROI are extracted. The features can be divided in three categories (Figure 9):

1. **Shape radiomics:** features of this category characterise the morphology, size and geometry of the ROI. 13 features within this category are calculated for every ROI: volume, surface area, surface area to volume ratio, sphericity, maximum 3D diameter, maximum 2D diameter (slice), maximum 2D diameter (column), maximum 2D diameter (row), major axis length, minor axis length, least axis length, elongation, flatness.

2. **Intensity radiomics:** features that quantify the CMR intensity. 18 features are extracted for every ROI: energy, total energy, entropy, minimum, 10th Percentile, 90th Percentile, maximum, mean, median, interquartile range, range, mean absolute deviation, robust mean absolute deviation, root mean squared, skewness, kurtosis, variance, uniformity.

3. **Texture radiomics:** The texture features quantify relations in intensities between neighbouring voxels. In total, 74 features are extracted using five different matrices: grey-level co-occurrence matrix (GLCM, 23 features), grey level run-length matrix (GLRLM, 16 features), grey-level size-zone matrix (GLSZM, 16 features), neighbouring grey tone difference matrix (NGTDM, 5 features), and grey-level dependence matrix (GLDM, 14 features).

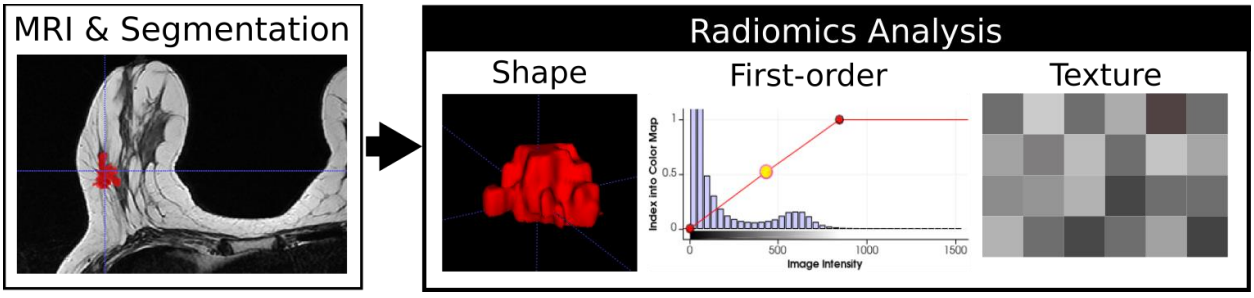

Figure 9: Overview of the three types of radiomics features extracted from Breast MRI images and corresponding lesion segmentation masks.

2.2 Tool overview

Figure 10 shows the front page of the radiomics analysis tool. The produced output file will be used in the future as input to the Machine Learning Toolbox to perform diagnosis and identify the most informative features for the task at hand.

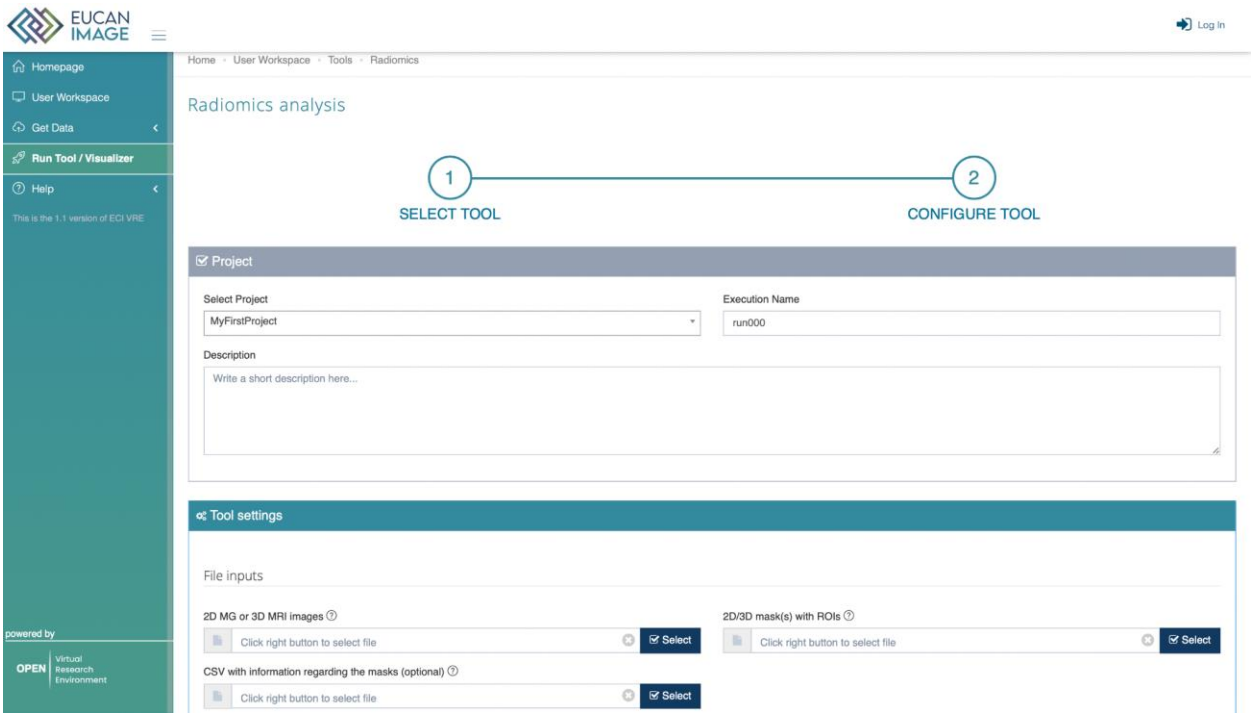

Figure 10: Main page and configuration settings for the radiomics analysis tool. The input images and mask() must be defined prior to its execution.

2.2.1 Input

The tool accepts as input:

1. Mammograms in 2D or Magnetic Resonance Images in 3D should be provided in NIFTI format (.nii or .nii.gz).
2. Segmentation mask in NIFTI format: 2D or 3D depending on whether the input images are MG or MRI, respectively. One segmentation per image should be provided. Multiple lesions per image can be presented in a single NIFTI file. The segmentation mask filename must be the same as the corresponding MG or MRI images with the addition of the suffix "_label".
3. CSV with information regarding the masks (optional): a .csv file that contains the information regarding (1) the correspondence between ROI and the labels present in the segmentation mask in case of multiple lesions. The .csv should contain at least the following columns: (1) id: image filename, (2) label_x: mask label corresponding to a specific lesion. In case the csv file is not provided, and multiple lesions are present, random labels are assigned to each lesion.
4. Bin width (Optional): parameter for grey value discretization necessary used for radiomics calculation. If not specified by the user, the default value 25 (grey levels) is used. It is worth mentioning that an unlimited number of MG/MRI images and segmentations can be passed at the same time to the tool.

2.2.2 Output

The output of the Radiomics Tool is a .csv file. Each row corresponds to a participant for which the MG/MRI and ROI has been provided and each column to a radiomic feature. The feature names are formatted as feature_name_label, where "label" is the index of the lesion present in the image.

3 Radiomics Feature Selection: Application to Use-Case 7 on Treatment Response Prediction in Breast Cancer

Biopsy confirmed invasive breast cancer is further followed-up with neoadjuvant chemotherapy treatment. Response to chemotherapy can be: 1) no-response – no changes or tumour has grown; 2) partial response – tumour has shrunk; 3) complete response – no signs of tumour. Depending on the response, the clinicians decide whether the patient should be headed to a surgery or undergo another type of chemotherapy. Response prediction to neoadjuvant chemotherapy is crucial and it allows planning treatment regimens for the patients to reduce the dangers and side-effects of chemotherapy.

In this use-case study, a dataset of 100 patients was collected and the tumours were semi-automatically segmented by undergraduate students after the lesions were localised by an expert radiologist. The study was conducted to evaluate the possibility of applying ML with radiomics features extracted from pre-chemotherapy MRI images.

The radiomics extraction was done using the tool described in this deliverable and a CSV file was downloaded from the EuCanImage VRE as shown in Figure 11.

	A	B	C	D	E	F	G	H	
1		original_shape_Elongation	original_shape_Flatness	original_shape_LeastAxisLength	original_shape_MajorAxisLength	original_shape_Maximum2DDiameterColumn	original_shape_Maximum2DDiameterRow	original_shape_Maximum2DDiameterSlice	original_shape
2	0	0.500136683184919	0.428288138200065	18.0422437299406	42.1264147210927	44.8035712862267	43.3663463990224	26.5224433263604	
3	1	0.870482779333484	0.705816339579922	25.0683786909561	35.5168579773542	41.37825515896	43.2832531124914	41.4477984940093	
4	2	0.788945900268043	0.601141245714678	13.5272291581097	22.5025803079402	63.0114275350115	57.1135710667789	26.2427132743548	
5	3	0.878567910628547	0.588144404405742	12.0294240647031	20.4531811823621	23.6676995079792	25.7650926643007	22.2	
6	4	0.949467942504958	0.695695171370447	15.2652158928536	21.9423916120946	24.4899979583503	26.2975284009734	30.6528954586675	
7	5	0.786689802612711	0.427662572138055	14.0133831858103	32.7673827423144	32.4499614791759	39.1152144312159	40.6452949306559	
8	6	0.773588343521738	0.464371706841528	21.8057652268055	46.9575663321947	48.674017709657	51.9649882132191	44.5457068638494	
9	7	0.895339775968005	0.801859260901486	20.4267552917069	25.4742400415034	32.538592471095	33.2770190972689	28.5047364485273	
10	8	0.737417982006671	0.659602004968174	15.7235465796517	23.8379302385692	25.1284699096463	25.8835855321476	26.3590591637866	
11	9	0.752142834771116	0.624860048584791	15.9564247628384	25.5359976989683	26.3180546393536	30	26.6308092254066	

Figure 11: Radiomics features extracted from 100-subject MRI dataset.

The features were normalised, and a number of feature selection algorithms were tested such as k-best using chi^2 and ANOVA F-value, variance based selection, sequential and recursive feature selection algorithms. The best performing setup in an 8-fold cross-validation was using the k-best feature selection algorithm. Among selected, only 4 were shape-based and the remaining were texture-based features. More improvements could be achieved by correcting the manual segmentation masks further with expert radiologists.

The results are shown in Figure 12 where the Areas Under Receiver Operating Characteristic Curve (AUC-ROC) of 0.80 and 0.81 were achieved using RBF-SVM and AdaBoost with Random Forest classifiers.

The preliminary results will pave the way towards building more robust prediction and decision support systems that will be integrated into the VRE of EuCanImage.

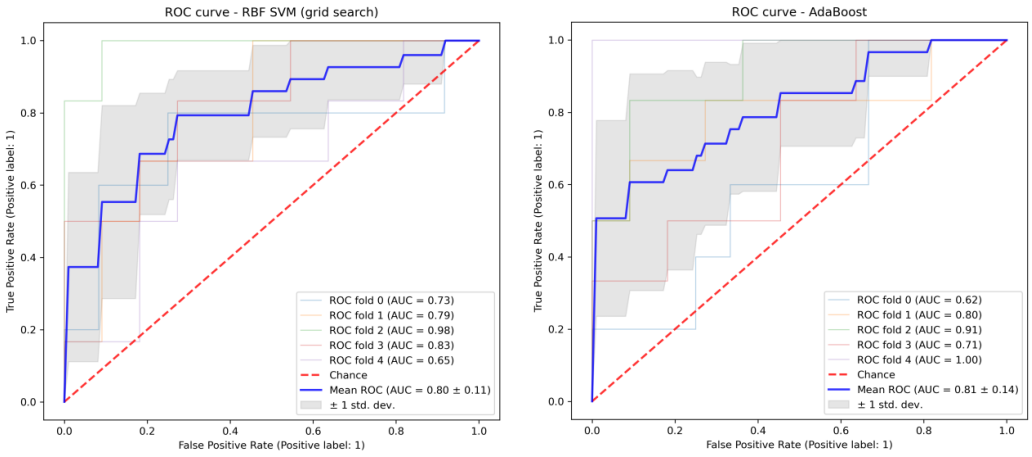

Figure 12: Preliminary results for the treatment response prediction tool for breast cancer using MRI images.

4 Next steps

Currently, EuCanImage is actively undertaking the annotation process for Liver, Colorectal, and Breast cancer use-cases. With more available data from the clinical partners, our next step will be to prepare a complete ML Toolbox that will enable researchers and doctors to easily build pipelines without prior expert knowledge. The ML toolbox will be built on top of the existing WORC framework [9] and integrated in the VRE. Two types of configurations will be available: 1) Preset tool – default pre-tested parameters; and 2) ML Mixer deck – allows users to build a pipeline based on image processing, radiomics features, feature pre-processing, feature selection, and ML model. The ML Mixer deck allows to run an arbitrary sequence of operations and it will initiate a search space for each step automatically to estimate the best parameters.

Overall, the workflow of the ML Toolbox will consist of a search space depicted in Figure 13. The search space consists of various sequential sets of algorithms, where each algorithm may include various hyperparameters, as indicated by the leaves in the trees. An example of a workflow, i.e., a specific combination of algorithms and parameters, is indicated by the grey nodes. The search workflow will be integrated in the VRE ML Toolbox. Figure 13 shows a prototype interface of the VRE tool which is under development and the planned report is due September 30, 2022. The progress that has been made on Radiomics feature extraction will aid greatly towards the successful completion of the ML Toolbox.

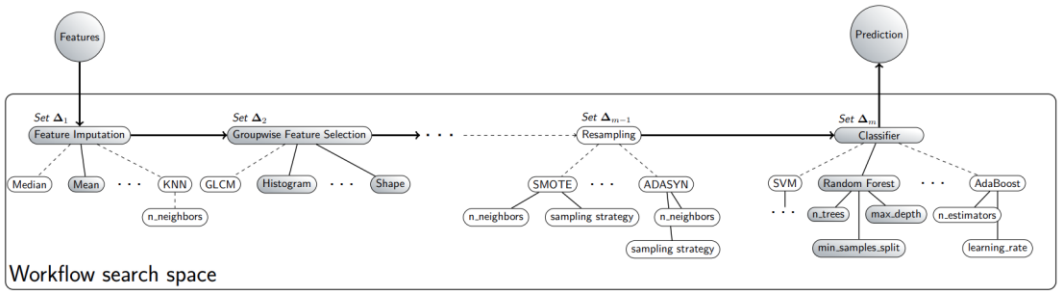

Figure 13: Schematic overview of the workflow search space in the ML Toolbox. Abbreviations: AdaBoost: adaptive boosting; ADASYN: adaptive synthetic sampling; KNN: k-nearest neighbour; GLCM: grey level co-occurrence matrix; SMOTE: synthetic minority oversampling technique; SVM: support vector machine. Figure courtesy of Starmans et al [9], EuCanImage partner.

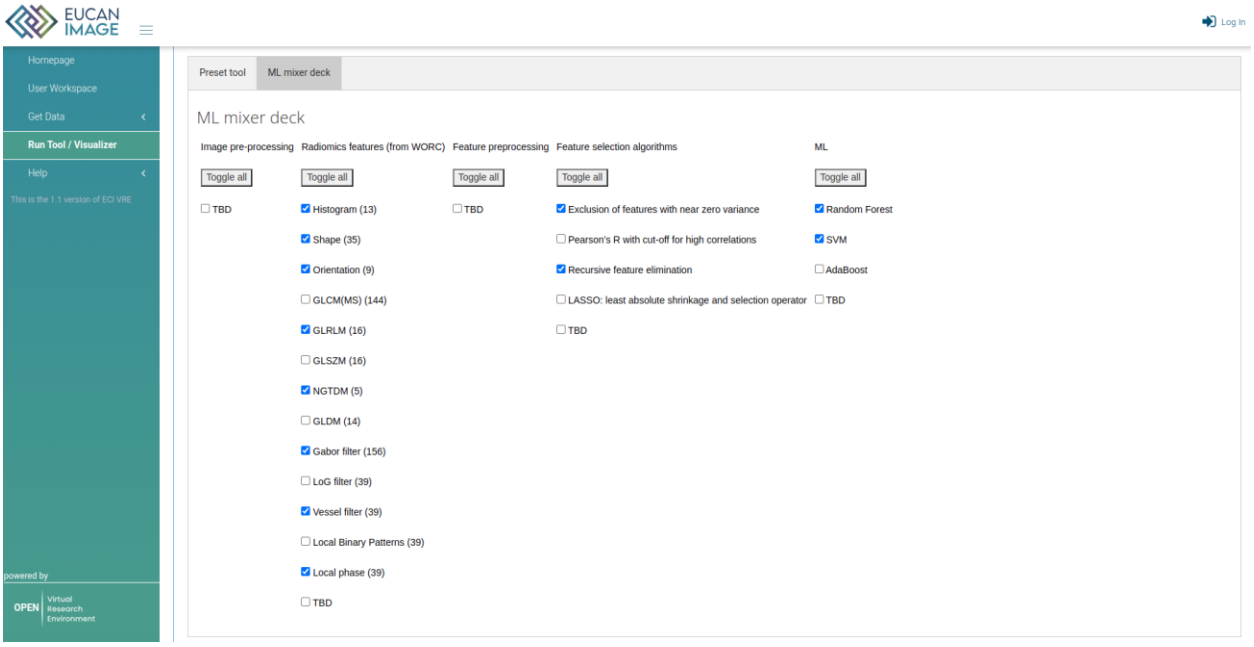

Figure 14: Machine Learning Toolbox Prototype (Under development and will be presented in Deliverable 5.3 which is due September 30, 2022). Two tabs will be available: 1) Preset tool – default pre-tested parameters; and 2) ML Mixer deck – allows users to build a pipeline based on image processing, radiomics features, feature pre-processing, feature selection, and ML model.

5 References

[1] Garrucho L, Kushibar K, Jouide S, Diaz O, Igual L, Lekadir K. Domain generalization in deep learning-based mass detection in mammography: A large-scale multi-center study. arXiv preprint arXiv:2201.11620. 2022 Jan 27.

[2] Nyúl, L.G., Udupa, J.K. and Zhang, X., 2000. New variants of a method of MRI scale standardization. *IEEE transactions on medical imaging*, *19*(2), pp.143-150.

[3] DeVries T, Taylor GW. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552. 2017 Aug 15.

[4] Xu Z, Liu D, Yang J, Raffel C, Niethammer M. Robust and generalizable visual representation learning via random convolutions. arXiv preprint arXiv:2007.13003. 2020 Jul 25.

[5] Zhou K, Yang Y, Qiao Y, Xiang T. Domain generalization with mixstyle. arXiv preprint arXiv:2104.02008. 2021 Apr 5.

[6] Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159. 2020 Oct 8.

[7] Siviengphanom S, Gandomkar Z, Lewis SJ, Brennan PC. Mammography-based Radiomics in Breast Cancer: A Scoping Review of Current Knowledge and Future Needs. Academic Radiology. 2021 Nov 16.

[8] Ye DM, Wang HT, Yu T. The application of radiomics in breast MRI: a review. Technology in Cancer Research & Treatment. 2020 Apr 28;19:1533033820916191.

[9] Starmans M, van der Voort SR, Phil T, Timbergen MJ, Vos M, Padmos GA, Kessels W, Hanff D, Grunhagen DJ, Verhoef C, Sleijfer S. Reproducible radiomics through automated machine learning validated on twelve clinical applications. arXiv preprint arXiv:2108.08618. 2021 Aug 19.