Pilot 8: Breast Cancer

1. Key Information

1.1 Involved Partners

- Institut Curie
- Teknologian tutkimuskeskus VTT

- IBM

1.2 Involved Countries

- France
- Israel
- Finland

1.3 Keywords

- Clinical decision support
- Predict patient response to breast cancer treatments
- Predict treatment effectiveness
- Precision medicine

1.4 Task Description

The pilot is a retrospective study that analyses mammograms, ultrasound and MRI images along with structured clinical data and information extracted from pathology reports to automatically predict patient response to breast cancer treatments, specifically neoadjuvant treatments. The data includes patients that obtained neoadjuvant chemotherapy (NAC) since 2012. The data is made available offline to the processing collaborators in IBM and VTT, through a VPN to access a local server at CUR. In summary, given images and clinical information, the models predict the probability of various outcomes for each patient who received neoadjuvant treatment. These models allow for evaluating the ability to make personalized treatment decisions rather than following global population guidelines and allow for assessing the economic effect of such protocols.

The overall methodology and pilot architecture are depicted in Figures 8a and 8b. To comply with regulations as GDPR, we use a model-to-data paradigm where all the data remains at Institut Curie infrastructure. All computations are applied on a strong GPU enabled server that resides in Curie, and various docker containers and pipelines of analytics models are transferred to the server and executed there. The overall flow is as follows: the anonymized imaging and clinical data are transferred from Curie clinical repositories to the pilot server hosted within the institute infrastructure. Training and inference pipelines utilize the data on the pilot server and produce analytics results. The analytics results are stored in a repository and an application is used to visualize those analytics results.

The pilot includes multiple pipelines. For example, one of the pipelines published in SPIE Medical Imaging 2020 predicts pathologic complete response using clinical and mammography (MG) imaging data. Another pipeline, published in PRIME-MICCAI 2020 predicts relapse using clinical and multiparametric Magnetic Resonance Imaging (MRI) data.

2. Building Blocks

2.1 Architecture

2.1.1 System Architecture

The amount of medical imaging is constantly growing in the number of images, in the size of each image and in the amount of information expressed in each image. At the same time, the field of computer vision is rapidly progressing incorporating new advanced technologies from image processing, machine learning, deep neural networks and general artificial intelligence techniques. Our aim is to create a scalable and generic architecture that can support all this growing data and new methodologies. We will utilize this architecture to research algorithms that automatically predict patient response to neoadjuvant breast cancer treatments.

The research and development cycle of Deep Learning (DL) and other new medical imaging algorithms includes several steps that our architecture needs to support. The research typically occurs in phases where we first use smaller dataset and later on scale to larger datasets. The cycle steps are (see Figure 8a below):

- Data preparation in the healthcare premise
    - Define the cohort according to the requested task and anonymize it
    - Annotate cohort to establish the ground truth
- Data retrieval to research platform
    - Retrieve annotated anonymized data and make it accessible for research
    - Create data splits for training, testing and hold-out.
- Deep learning model creation (iteratively)
- Test models, utilize relevant quantitative measures and tune hyperparameters. Then select the most accurate model or ensemble of models
- Model deployment
- Create model as a service and evaluate its accuracy on hold-out data split
- Deploy in a scalable platform and run service on real-world data and new scenarios

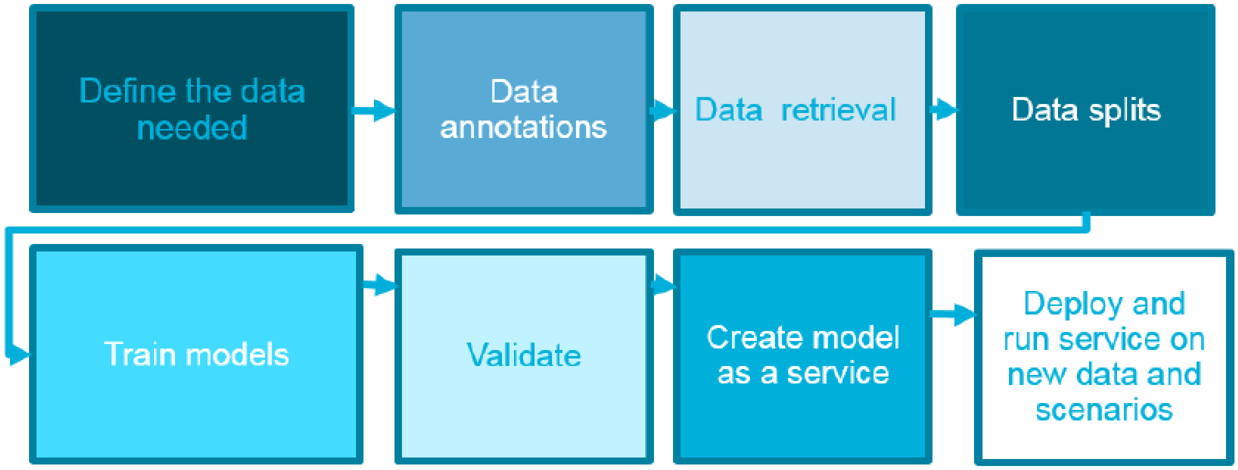


Fig 8a: Deep Learning Methodology

To support the above steps performed by the various pilot partners, our architecture needs to enable the integration of polyglot algorithms that may be written in different programming languages (Python, Java, C) and use different deep learning or other frameworks. Moreover, the DL algorithms are compute intensive and have special requirements that need to be considered and optimized such as GPUs, large memory usage, and large disk capacity.

The overall pilot architecture is depicted in the figure below. A strong server with GPUs located in Curie data center will be used to run docker containers with various diverse polyglot analytics modules. The models may be also ensembled to intelligently combine algorithms and improve overall performance. IBM will use a generic Biomedical Framework that creates configurable reusable pipelines and exposes them as REST microservices. The framework also enables transparent run of pipelines on a SPARK cluster where workers run in parallel on the same server and each worker runs on a separate GPU. It is doing that by automatic translation from a descriptive pipeline flow to efficient SPARK application that can perform multi-modal analytics and utilize analytics modules written in various programming languages.

The anonymized imaging and clinical data are exported from Curie ConSoRe repository to the pilot server and to enrich our data, we also add open data of similar characteristics. Training and inference pipelines utilize that data and produce analytics results that may be evaluated against the ground truth. The analytics results are stored in a repository and a demo app is used to visualize those analytics results.

The proposed architecture is scalable and can easily grow from one node to as much as needed without changing the algorithm code. The architecture enables scale in all axis: x-axis (containers), y-axis (function partition), z-axis (data partition). By using the Apache SPARK open source, we get a fault-tolerant, distributed backend for robustly analyzing large datasets in a scale-out cluster.

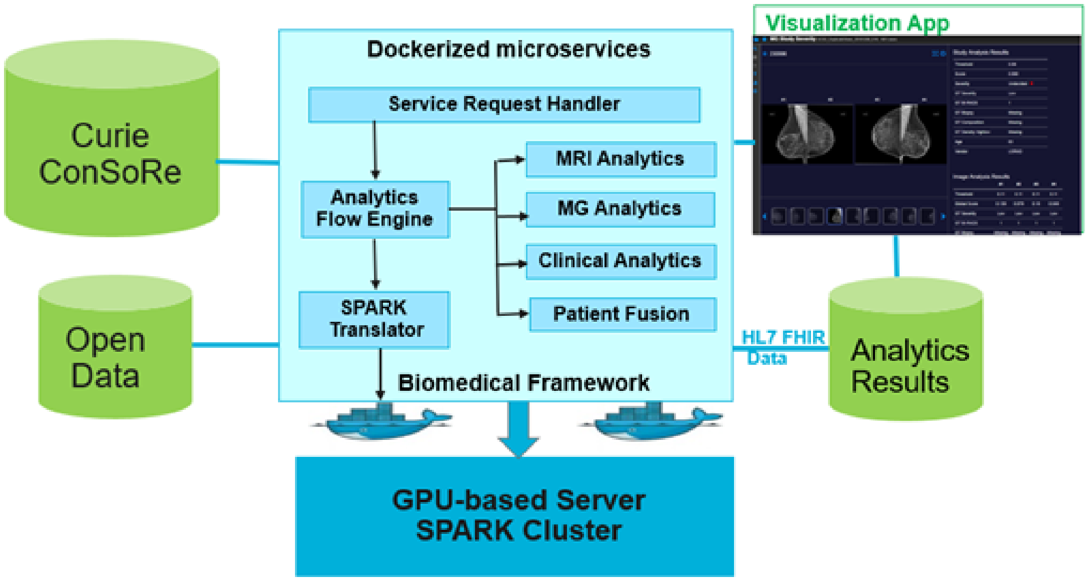


Fig 8b: Pilot Architecture

2.1.2 Data Flow & Interoperability of services

The Breast Cancer pilot analyses mammograms, ultrasound and MRI images along with structured clinical data to automatically predict patient outcomes in neoadjuvant treatments. We develop models to predict for each patient the probability of various possible outcomes including pathologic complete response, relapse, five-year recurrence. These models, pending clinical trial validation, will allow evaluating the ability to make personalized treatment decisions rather than following global population guidelines and will allow assessing the economic effect of such protocols.

Institute Curie's clinical data pipeline extracts data from several sources and anonymizes it, creating several csv files with relevant clinical data as depicted in Figure 8c. These files are then stored in two servers at the Institut Curie, which can be accessed by VTT and IBM. Similarly, three kinds of images are available; mammography, Magnetic Resonance Images (MRIs) and UltraSound (US) images. All the available images have been anonymized and made available in the pilot servers, together with the clinical data.

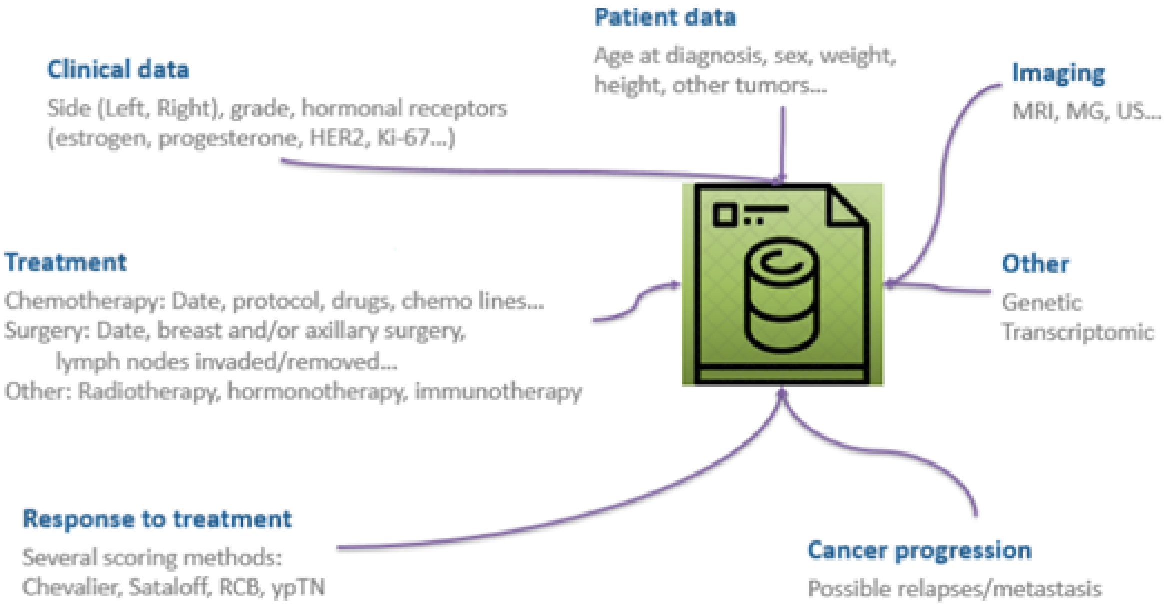

Patient data
Age at diagnosis, sex, weight, height, other tumors...
Clinical data
Side (Left, Right), grade, hormonal receptors (estrogen, progesterone, HER2, Ki-67...)
Imaging
MRI, MG, US...
Treatment
Chemotherapy: Date, protocol, drugs, chemo lines...
Surgery: Date, breast and/or axillary surgery, lymph nodes invaded/removed...
Other: Radiotherapy, hormonotherapy, immunotherapy
Other
Genetic
Transcriptomic
Response to treatment
Several scoring methods:
Chevalier, Sataloff, RCB, ypTN
Cancer progression
Possible relapses/metastasis

Fig 8c: Data collection and flow

Various models and pipelines were developed on top of the collected data. IBM pipelines used the clinical data, the MG data, and the DCE-MRI subtraction data to create individual models. Afterwards, we created an ensemble model on top of all the individual models.

IBM pipelines for MRI use annotated DCE-MRI subtraction volumes. A DCE-MRI scan of a patient with breast cancer includes multiple volumes. The volumes are taken before a contrast agent is injected, and at several intervals after the injection. For our analysis, a digital subtraction of the volume acquired after injection of the contrast agent and the baseline volume acquired before the injection. We chose to use the subtraction volumes because this type of imaging is used by radiologists for medical diagnosis and was likely to contain the information relevant for our analysis. The MRI subtraction slices are used as input to a deep neural network (DNN). The slices are first preprocessed, then each slice is transformed via 2D CNN, and then the collected features from all slices are transformed via a 3D CNN.

VTT image processing pipeline creates first 3D or 4D volumes from the DICOM slices of the MRI images. 3D volumes are created from static imaging and 4D volumes from MRI time-series. After creating the volumes, VTT pipeline generates binary breast masks using Dixon MRI series water-only and fat-only volumes and additionally apparent diffusion coefficient (ADC) volume generated from diffusion-weighted MRI data. The generated MRI binary mask is cut to half and the lesion breast mask is again segmented into lesions and other breast tissue. The lesion segmentation is then used to calculate explainable features for IBM DNN network.

2.1.3 Necessary Hardware

We have two servers in Curie for the pilot. Each server has the following configuration:

CPU Cores - 12 cores
RAM - 128 GB of RAM
Storage - 8 TB HDD

In addition, one of the servers has 2 GPUs of type Nvidia Tesla V100.

2.1.4 Software Components

IBM components include various models developed with open source frameworks for deep learning including Tensorflow, Keras, PyTorch. The models are wrapped with IBM Biomedical Framework, a platform to create configurable reusable pipelines and expose them as micro-services on-premise or in-the-cloud. One important feature of the Biomedical Framework is the enablement of running pipelines of AI models on distributed environments such as SPARK cluster and doing that transparently to the algorithm developer. Given a description of the pipeline with the algorithms dependency graph, the Biomedical Framework automatically translates the pipeline descriptor to an efficient SPARK application. The resulting application is efficient in the sense that it minimizes the time from the beginning of the first algorithm until the completion of the final algorithm in the pipeline.
The results of the pipelines are visualized in IBM Experiments Viewer that is a Web application to explore results of analytics pipelines (specifically deep learning) and their evaluation on various datasets.

In the VTT pipeline, the MRI volumes are constructed with dcm2niix command line software from MRIcron application package (https://www.nitrc.org/projects/mricron) and converted to nhdr format (http://teem.sourceforge.net/nrrd/format.html) with SimpleITK python package (https://simpleitk.org/).The segmentation and feature calculations were developed as custom Matlab functions and packaged as custom Python package, which was transferred to Curie secure environment and ran with standalone Matlab Runtime (https://se.mathworks.com/products/compiler/matlab-runtime.html) through Python scripts. Volumetric segmentation results were stored as compressed nhdr files. Numerical volumetric features were exported as csv files for IBM network.

2.3 Data Processing

2.3.1 Processing of large structured / unstructured data sources

2.3.1.1 Data Sources

All our medical records at Institut Curie (IC) are entirely in an electronic format. Most of the data has been entered in a semi-structured way, namely it's a mix of structured data and free text. From these records, a cohort of around ~1700 patients have been identified matching the required criterias:

- Women with breast cancer who have received neoadjuvant chemotherapy (NAC) since 2012.
- Excluded multifocal and bilateral cases. Also excluded patients with skin invasive or inflammatory tumours.
- Excluded patients who have relapsed from a previous tumour event.

In the generation of clinical data, a set of potentially relevant clinical metadata has been identified by medical doctors at IC. The metadata dictionary describes the identified clinical data and provides a better understanding of them. The set of clinical data includes:

- Patient information: age at diagnosis, weight and height.
- Tumour properties: side of the tumour (left or right breast), grade of the tumour, percentage of stromal tumour-infiltrating lymphocytes, hormonal receptors (estrogen, progesterone, HER2).
- Neoadjuvant treatment: timing of chemotherapy, whether it includes targeted therapy, properties of the surgery including response to treatment, radiotherapy events.
- Evolution after treatment: relapse and metastatic events.
- A binary classification (manually verified) regarding the complete response to treatment.

Clinical data was extracted from a multi-purpose SQL repository. This repository has been developed at Curie in order to aggregate information coming from all different sources (software used by doctors, medical records, handwritten events, etc). The idea is to have access to all existing information in Curie for every patient that is treated with NAC since 2012 (personal data, tumour events, treatments, response to treatments, relapses, etc). This repository has already been used for other projects and has proven to be extremely valuable in terms of quantity and quality of information.

For the Breast Cancer Pilot, three kinds of images are available at IC; mammography, Magnetic Resonance Images (MRIs) and UltraSound (US) images. IC does not have every kind of image for every patient at every stage of the cancer treatment. Images available at all stages are only available for a handful of patients. Thus, images are very valuable not only because of the amount of information they contain, but also because of their limited supply. All the images are currently being anonymized, a time-consuming process as we are validating the full anonymization at every stage.

Making the data available to VTT and IBM was a two-step process, in which the data never left Curie premises.

- First, a subset of several patients were identified maximizing the heterogeneity of their health records. The clinical data belonging to these patients was manually curated and delivered with the corresponding images all fully anonymized. The complete collection of clinical data and images was shared with VTT and IBM.
- Second, a similar process of manual curation plus image anonymization was carried out for the rest of the patients, and their data was shared with VTT and IBM and hosted in a server with GPUs.

Data Source	**Description**	**Acquisition**	**Characteristic (Size, Patients, Years, Origin/Region)**
Clinical - IC internal repositories	We have a collection of databases which we query to extract different kinds of data; about the patient, about the tumor, about the treatment and about the response to treatment and outcomes. We anonymise the data and aggregate it per patient, generating a list of relevant clinical features per patient.		We extract around 50 features for about 2200 patients, diagnosed between 2012 and 2020.
Imaging - IC internal repositories	Similar to the PACS, we have an internal server hosting the images of our patients, which we query to extract the relevant images and anonymize them.		MRI, MG, US for about 800 patients

multiple sources	**integration to data warehouse**	**data access**	**data stored in cloud**	**multi-party architecture**	**secure environment**	**transform raw / unstructured data**
yes	yes	access via VPN, process data only in hospital	no	no	yes	yes

		and anonymized				

2.3.1.2 De-Identification and anonymisation

Each type of image has a different anonymization procedure. MRI and mammography are DICOM images which contain both pixel data and a list of metadata tags. Some of these tags are nominative, like name or date of birth of the patient, so we need to create a complete list of existing tags and remove those who could be used to find the identity of the patient. A priori there is no need to modify the pixel data of the DICOM images. Regarding US images, they sometimes contain the name of the patient somewhere at the edge of the figure. We have a procedure to crop the images and produce new ones where the name of the patient is not visible.

For the clinical data, we anonymized the patient ID as well as all the dates, changing them to relative days since the patient was born.
The same anonymized patient ID was used for the clinical data and the imaging data, so we could correlate them.

2.3.1.3 Acquisition

Image data was provided in bulk, with some structured clinical data embedded in the DICOM image header.

Additional clinical information was extracted from the EHR systems in CUR and restructured into patient level flat records. This included the index date of treatment initialization, some summary clinical and demographic features prior to diagnosis, tumour properties, NACT treatment indications, surgery properties, and the outcome of the treatment (e.g. pathologic complete response, relapse, five-year recurrence).

2.3.1.4 Cleansing

We will exclude from the data cases with multifocal tumours, bilateral cancer, skin invasive cancer or inflammatory tumours. We will also exclude patients who have relapsed from a previous disease.

Cleaning and filtering of image data included removal of images with low resolution, images with obstructions, and images with previous surgical indications. We preprocessed the images to include just the relevant side of the breast and relevant part of the tumour.

Clinical data was examined to identify extreme outliers, and to identify features with extreme bias between treatment groups. Patients who were extreme outliers were excluded from further analysis, and features that showed extreme bias were further analysed to check if the bias could be corrected. We also handled censored patients and included them in train data but not in validation or test.

The clinical outcomes were analysed for inconsistencies between the various data sources. The association between features and outcomes were analysed to see the baseline predictive power available in the data.

2.3.1.5 Data Integration

The various data sources are integrated at the patient level.

Each image file is identified by an anonymized patient id, and this ID is used to identify the rest of the clinical and outcome features.

The analysis itself creates algorithms based on all the inputs in the relevant stratified data. The algorithms may be polyglot, namely written in different programming languages (Python, Java, C) and use different deep learning or other frameworks (Tensorflow, Pytorch).

2.3.1 Multi-velocity processing of heterogeneous data streams

Does not apply.

2.3.5 Complex real-time event detection

Does not apply.

2.3.5.1 Notifications

2.3.5.2 Situations of Interest

2.3.5.3 Event Processing

2.3.5.4 Event Sources

2.3.5.5 Evaluation

2.4 AI Components

2.4.1 Deep learning for multilingual NLP and image analytics

2.4.1.1 Natural Language Processing

Does not apply

2.4.1.1.1 Evaluation

2.4.1.2 Image Processing

Type	How will IA support your pilot?	How will IA help you to reduce costs?
Medical imaging	Medical imaging is a non-intrusive method to get valuable personalized information about the patient's condition. Specifically, in breast cancer, there are several modalities: magnetic resonance, mammography, and ultrasound to capture the patient's condition, and performing image analytics on this data enables us to give personalized and more effective treatment.	By using image analytics to predict response to neoadjuvant chemotherapy (NAC) treatment, we can influence the right treatment for the patient. This can save the costs of ineffective treatments as well as prolong patient's quality of life, making them productive and contributors to the EU economy for longer periods.

Do you require regulatory approval?	Which Image Processing tasks do you address?	Do you use public data repositories? If yes, which?	Describe your method in a few sentence.	Which training technique do you use?	Which pathologies are covered?	On which level does classification happen (volume, slice/img, or px-level)?	What is the level of detail of your GT-annotations (volume, slice/img, or px-level)?
No as we are doing a retrospective study	Multi-modal classification of imaging and clinical data including longitudinal mammography (MG) images, ultrasound (US) images and magnetic resonance (MR) images	Yes, ISP1 and UCSF NACT datasets from https://wiki.cancerimagingarchive.net	We'll use several methods and then ensemble them altogether. The methods include (1) deep neural networks to analyze images, (2) traditional feature extraction from images, (3) classical machine learning methods such as xgboost.	Train using convolutional neural networks (CNN)	Breast cancer	On volume for MRI and on slice/image for MG and US	Annotation is per volume level

2.4.1.2.1 Evaluation

For evaluation, we set aside a holdout cohort of 100 patients. From the rest of the patients, we created the train cohort. We split the train cohort into 5 folds with equally distributed positive and negative samples among folds. As the number of patients having imaging is smaller than the number of patients that have clinical data, we have different train cohorts for imaging and for clinical. In our split to folds, we make sure that the folds in both imaging train cohort and clinical train cohort are correlated, namely a patient remains in the same fold.

We performed cross-validation and computed the ROC, AUC with confidence interval, specificity at sensitivity for each fold, and the mean values across folds. We then evaluated our model on the holdout data and computed AUC with confidence interval, and specificity at several sensitivity operation points. We also examined the features of importance produced by our models, and used the SHAP algorithm for explainability.

2.4.2 Prediction Algorithms

2.4.2.1 Task

Prediction of (1) outcomes (pathologic complete response, relapse, five-year recurrence) of neoadjuvant chemotherapy (NAC) treatment and (2) prediction of cohorts for clinical trials towards next generation therapies.

2.4.2.2 Data, Data Modelling

We used two train cohorts in our experiments because some patients had only clinical data while other patients had clinical and imaging data. The first data subset is a large cohort of patients for clinical data evaluation. The clinical data included demographics such as age, weight, height, and tumor properties such as breast cancer histology, grade of the tumor, Ki67, and molecular subtypes based on estrogen, progesterone, and HER2. The second data subset was a smaller cohort who, in addition to the clinical data, also had imaging scans taken prior to NAC treatment. The small cohort is a subset of the larger cohort.

We annotated the MRI data. We annotated the most important subtraction volume in which the tumor appeared to be the brightest in terms of relative illumination. In the selected volume, we also annotated the significant slice in which the tumor was the largest.

2.4.2.3 Features

CNN

1. MRI images
2. MG images
3. US images
4. Features from classical image processing
5. Clinical data

XGBoost
1. Features from CNN
2. Clinical data

2.4.2.4 Model

We have developed several models as described below and created pipelines to combine them all together and produce the final output.

MRI deep learning model (IBM)

The MRI data includes a preprocessing stage before entering the convolutional neural network (CNN). The input to the CNN is the significant slice and the two pre and post adjacent slices (i.e., three slices in total) that are extracted from the selected MRI subtraction volume. The selected slices undergo a cropping and resizing process. Our data consisted of axial MRI volumes, which contain both sides of the breast. Hence, we cropped the image vertically and continued processing only the relevant side in which the tumor was located. Then, we cropped the image horizontally to exclude non-breast parts that appeared in the image. This process was done automatically using a sliding window, where we searched the most enhanced organs within the first slice in the MRI volume, and found a cut line above them that was used for our three selected slices. Each of the vertically and horizontally cropped slices was then resized to 512 x 256 pixels to bring them all to the same size. The last two steps of the pre-processing included rotating the slices, so the breast was facing in the same direction for all slices. We also under sampled the slices where there was overlap between slices in the volume.

Next, we use the preprocessed slices as input to our CNN, which is a modification of ResNet as a classifier. We specifically used ResNet18 formulation, but reduced the number of filters per layer to speed up training and avoid over-fitting. The original Resnet18 consists of blocks of convolutions, with residual connections between the blocks. Each convolution layer is followed by a batch normalization layer and ReLU activation. For our network, we used 7 residual blocks with [32, 64, 128, 128, 256, 256] filters per convolutional layer. This 2D-CNN model was applied simultaneously to the 3 slices, i.e., the same 2D-CNN model with the same weights was applied to each slice. Next, a 4D-tensor was used to aggregate features produced from the 3 input slices. Finally, a 3D convolution layer was applied, followed by a 3D average global pooling layer. The output of the pooling layer was treated as an embedding vector . On top of this embedding layer, we added a simple sigmoid-activated linear layer as an output layer. A detailed diagram of the CNN model is depicted in Figure 8d.

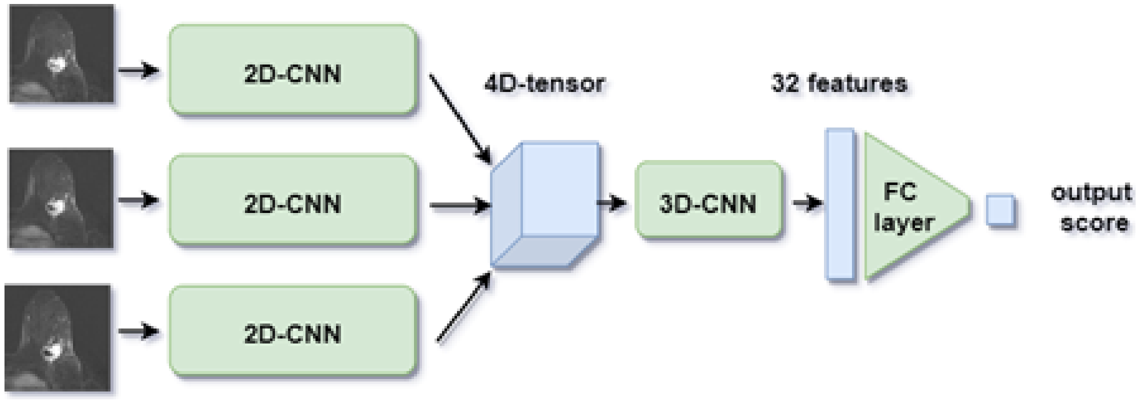


Fig 8d: MRI CNN architecture

MRI image processing features (VTT)

VTT’s original plan was to apply their existing interpretable models that are easily applicable to the breast MRI images although they were designed mainly for brain MRI. However, due to missing voxel level labeling in the dataset the original plan had to be abandoned and a new approach with unsupervised machine learning using fuzzy c-means (FCM) clustering was adopted for the segmentation task. The obtained segmentation masks were used to generate volumetric numerical features for IBM DL models. The VTT pipeline concept is visualized on figure below:

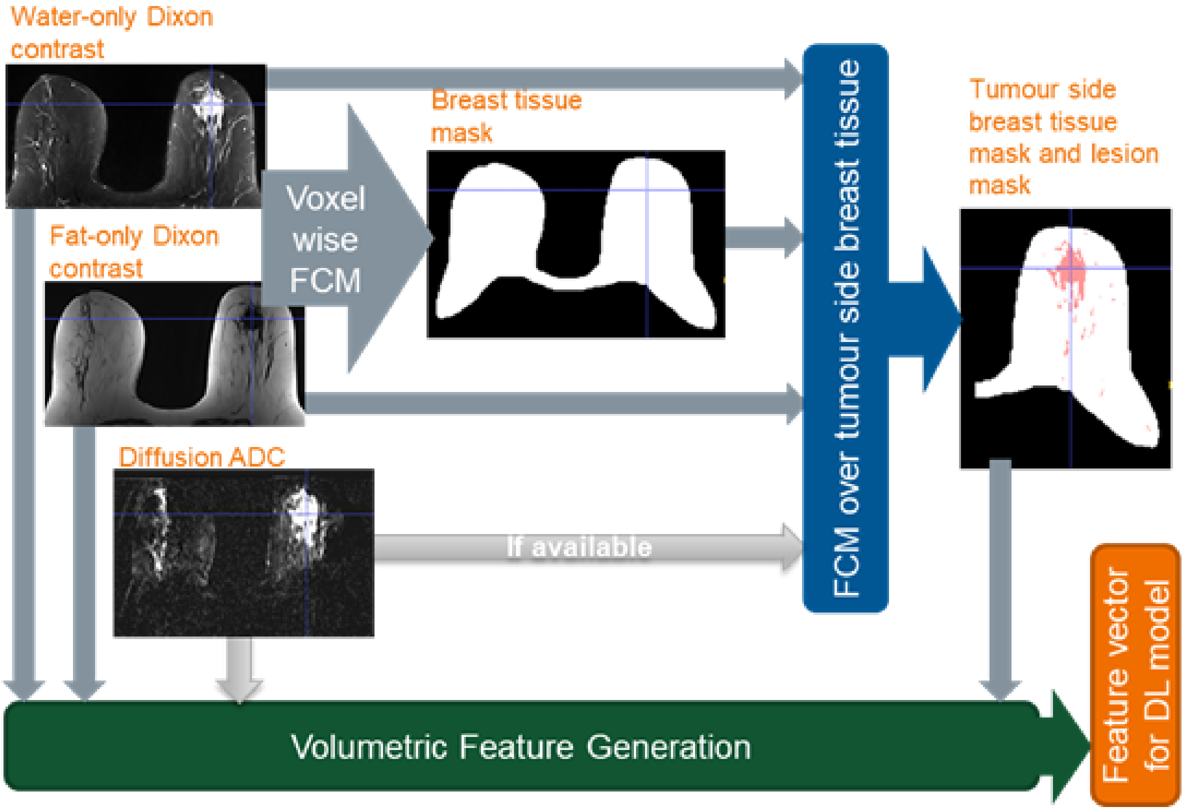

Water-only Dixon contrast
Fat-only Dixon contrast
Diffusion ADC
Voxel wise FCM
Breast tissue mask
If available
FCM over tumour side breast tissue
Tumour side breast tissue mask and lesion mask
Volumetric Feature Generation
Feature vector for DL model

Fig 8e: Image processing features

MG model (IBM)

We use a pre-trained deep learning (DL) model to get the tumor location. The model was trained on several thousands of MG images from IBM repository. We used that model to inference the MG data in Curie, and then compute radiomics texture features (Gabor, LBP, GLCM, wavelets) within tumor area and in the peritumoral area. Figure 8f below shows the network output predictions of tumor detection. On the left, there is the MG image from the Curie dataset with a detected contour around the tumor area. On the middle, the tumor patch is extracted from the detected area, and on the right, the tumor margins are extracted.

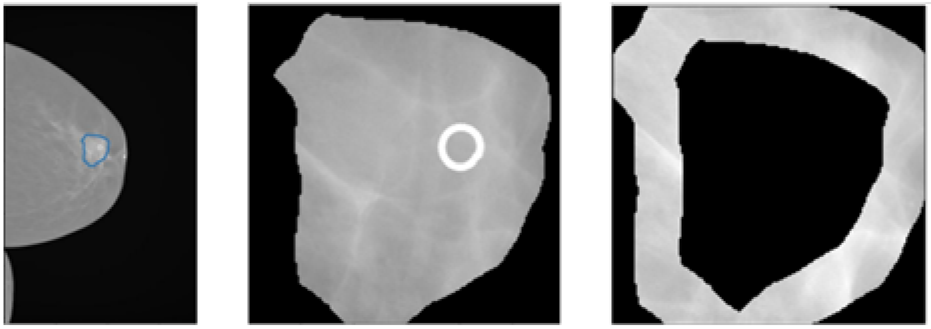


Fig 8f: MG processing

Clinical model (IBM)

We created our model using the clinical pre-treatment features per patient. The features have values in different ranges and some values are missing; thus, we preprocessed the data by applying a scaler that scales all features to the [0,1] range. An imputation process replaced missing values with the mean value. To select the best classifier for our task, we trained the data with three known machine learning algorithms: Random Forest, Logistic Regression, and XGBoost. We evaluated our model and examined the features of importance using the SHAP explainability algorithm.

Ensemble model (IBM)

The ensemble model depicted receives six scores per patient: three scores based on clinical data and three scores based on the imaging data. To improve generalization, we created multiple variations of each model where each different variation started its train from a different initialization. Thus, the three scores for clinical data are produced from three variations of the clinical model that differ in their training initialization. Likewise, the three scores for imaging data are produced from three variations of the imaging model that differ in their training initialization.

We examined several strategies for combining the models and evaluated the cross-validation AUC and specificity at sensitivity for each option. We first tried the stacking classifier, in which we trained a meta model on top of the six models' scores using the small cohort folds. We also tried several voting strategies. However, we found that the most effective strategy used the average value of all available scores per patient.

2.4.2.5 Evaluation

See 2.4.1.2.1

2.5 Security and privacy of data access and processing

See 2.3.1.2

2.5.1 Access Control

2.5.1.1 Authentication

2.5.1.2 Authorization

2.5.2 Data Protection

All the data is fully anonymized and doesn't leave Curie premises. Thus, there is no need to encrypt it.

2.5.2.1 Data at rest

2.5.2.2 Data in transit

2.5.3 Auditory and logs

All the data is fully anonymized and doesn’t leave Curie premises. Thus, there are no special audits or logs for our pilot.

2.5.3.1 System Auditory

2.5.3.2 Services Auditory

2.5.4 Privacy measurements

2.5.4.1 Data Privacy Impact Assessment (DPIA)

2.5.4.2 Legal/Ethical process

2.5.4.3 Processes for complying with the current legislation

2.6 Trustworthy AI

2.6.1 technology/user adoption and establishing trust

2.6.2 ethical principles

- respect for human authority
- prevention of harm
- fairness
- Explicability

We kept all above ethical principles and used known algorithms to support them when possible. In particular, we tried to provide explanations to our models, e.g. via the Shapley Additive Explanations (SHAP) algorithm for exaplanability. SHAP considers all possible combinations of features with and without that specific feature to evaluate its contribution to the prediction. The figure below is an example of how SHAP depicts the top 10 clinical features in descending order that had the most influence on predicting pCR. In the summary plot of SHAP, each point represents a single patient. The x-axis indicates the effect (either positive or negative) of the feature on the predicted score for the patient. The point's color represents the value of the features (red=high value, blue=low value).

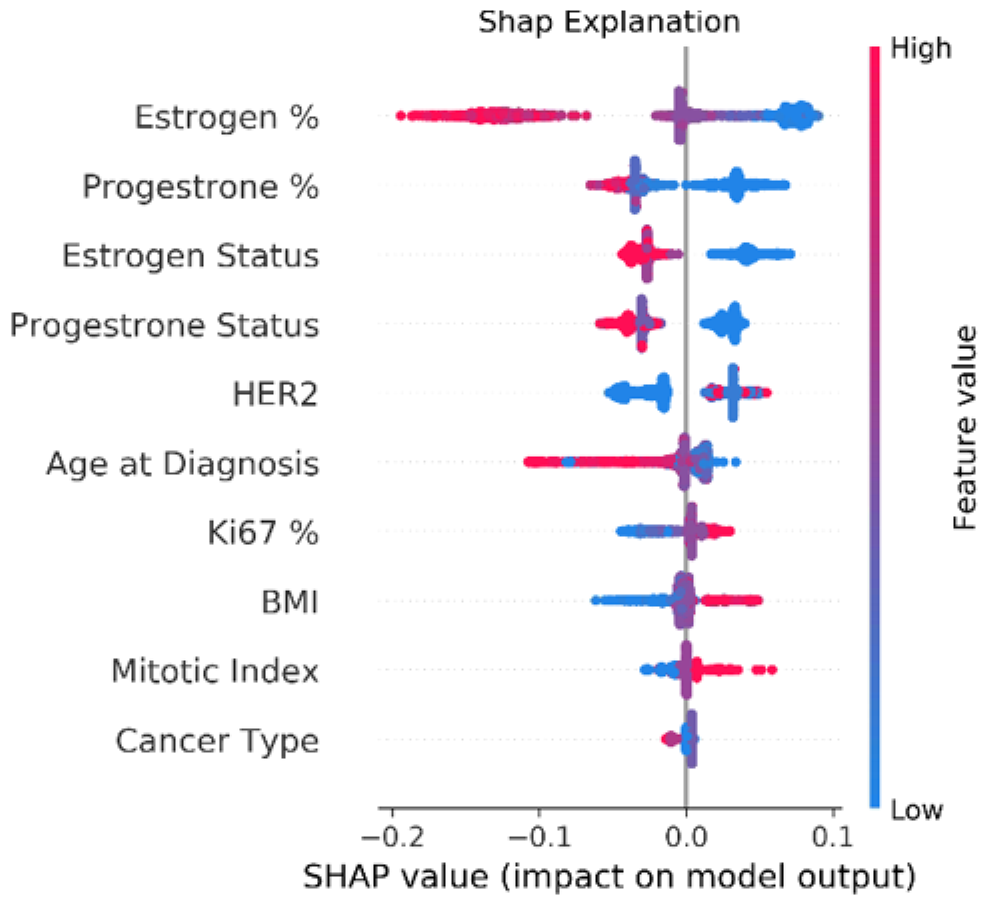


Fig 8g: Clinical features contribution to predict pCR

Moreover, the VTT MRI image processing pipeline aims to generate explainable features for the IBM DL network. The VTT features are by volumetric nature possible to visualize over MRI images, which makes it possible for human agents to overview and track the numerical feature content when necessary.

2.6.3 key requirements

- Human agency and oversight
- Technical Robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental well-being
- Accountability

We tried to consider all the above key requirements, and also all local and global regulations including:

- General data processing regulation (GDPR)
- Loi “Informatique et Libertés” (modified July 2019)
- Code de la santé publique
- EMA and FDA

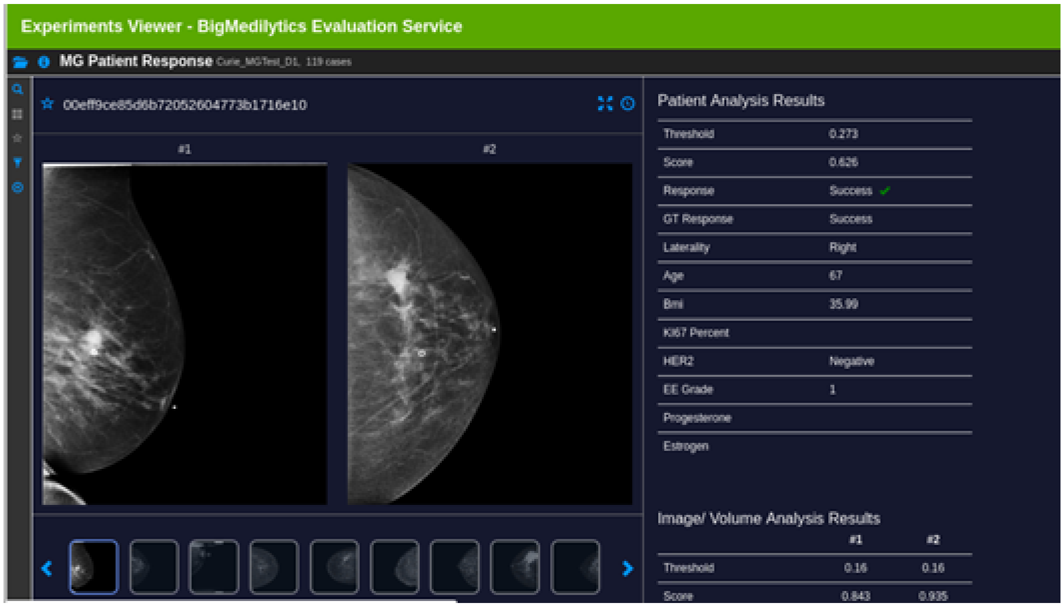
2.7 System-Interaction

2.7.1 Human-Machine Interface / GUI

We created a web application, the Experiments Viewer, that evaluates and visualizes the results of multimodal medical analytics pipelines, especially with deep learning imaging algorithms. More specifically, the Experiments Viewer functionality includes:

- For each patient, show the original imaging, the explanation imaging, the clinical data, the scores of all algorithms and the ground truth if available
- Search of selected patients in the experiment
- Filter experiment to focus on specific sub-group and examine it
- For each experiment, the evaluation per image-level and per patient-level. The evaluation includes confusion matrices, AUC graphs, specificity and sensitivity at different thresholds, sub-group analysis, etc.
- Visualize the data per image-level and per patient-level using Google Facets Overview and Google Facets Dive

The figure below shows the main screen of the Experiments Viewer and some of its graphs.


Experiments Viewer - BigMedilytics Evaluation Service
MG Patient Response Curie_MGTest_D1, 119 cases
00eff9ce85d6b72052604773b1716e10
#1
#2

Patient Analysis Results

Threshold	0.273
Score	0.626
Response	Success ✓
GT Response	Success
Laterality	Right
Age	67
Bmi	35.99
Ki67 Percent	
HER2	Negative
EE Grade	1
Progesterone	
Estrogen	

Image/ Volume Analysis Results

	#1	#2
Threshold	0.16	0.16
Score	0.843	0.935

Fig 8h: Experiments Viewer

2.7.2 Education

We created papers and demos to publish our results and review with the community.

3. Learnings

3.1 Challenges & Barriers

- Architecture
- Processing of large structured / unstructured data sources
- Multi-velocity processing of heterogeneous data streams
- Complex real-time event detection
- Natural Language Processing
- Image Processing
- Prediction Algorithms
- Security and privacy of data access and processing
- Trustworthy AI
- System-Interaction

While our MRI data for predicting relapse after NAC treatment is one of the largest compared to those reported in prior art, it is relatively small for deep learning networks. Moreover, MRI has no standardized protocol for scan acquisition and high variance of image resolution, voxel size, and image contrast dynamics. We selected special MRI preprocessing and neural networks to adjust for these limitations, and the major contribution of this modality to our prediction is clear. Yet, to get robust models that are not sensitive to fold partitions and generalize better, we need to retrain our models on much larger datasets.

VTT faced challenges in image processing for two main reasons. The first challenge was that all the imaging data to be used in the analysis was real world patient data collected in the past years within clinical practice. The availability of imaging modalities and sequences, quality of the data and even the naming and vocabulary of the studies varied a lot from patient to patient. As there was no table of contents available but only a pool of mixed data, the first challenge at the beginning of the project was to understand what kind of (MRI) images exist in the dataset in sufficient amounts. We also had to investigate what are the used protocols and how many subjects are having consistent images.

The second challenge for VTT original plans was that there were no voxel level annotations available on any of the images or modalities and there were no resources on project to generate these. This led VTT to a situation that we could not use our existing solutions and we had to modify our approach to unsupervised machine learning (ML) instead of supervised ML. This approach was still only a partial solution for the challenge as without true voxel labels we could confirm only visually the correctness of the segmentation.

3.2 Lessons Learned

- Architecture

- Processing of large structured / unstructured data sources

- Multi-velocity processing of heterogeneous data streams

- Complex real-time event detection

- Natural Language Processing

- Image Processing

  It is important that when medical images are used in the project, the clinical specialists for medical imaging are available for the project. These can be radiologists or other medical operators who are able to detect and explain the targeted phenomenon in the medical images. Moreover, it is important that image datasets are generated in close collaboration with clinicians who have been involved to collect the data originally.

- Prediction Algorithms

We used deep learning and image processing algorithms to analyze our mpMRI data and classical machine learning algorithms to analyze the clinical data. Using two branches enabled us to use the best method per modality and utilize the maximum available data for each data type. This approach improved our results.

High sensitivity was important in our problem setting since we wanted almost all patients that encountered recurrence to be correctly classified by our model and enable treatment options to be reassessed in advance. It is also important to know the specificity in these high sensitivity operation points. Adding the MRI modality enabled us to improve the specificity at high sensitivity operation points.

- Security and privacy of data access and processing

- Trustworthy AI

Explainability should be done in collaboration with the doctors to make sure the visualized information is valid and valuable for them.

- System-Interaction

3.3 Main (quantifiable) achievements

Here are some of our main achievements:

- Created fully anonymized cohort of 1738 patients treated with NAC; from which almost 600 patients have imaging prior to treatment start (MRI, MG, US)
- Created annotations for the MRI data
- Created MG, clinical and ensemble models to predict pathologic complete response (pCR)
- Created MRI, clinical and ensemble models to predict relapse
- Created multiparametric MRI, clinical and ensemble models to predict five-year recurrence
- Built a system in Curie that runs pipelines based on the Biomedical Framework
- Enhanced the Experiments Viewer to visualize our cohort and tasks

4. Output

4.1 Papers

- Paper: "Radiomics for predicting response to neoadjuvant chemotherapy treatment in breast cancer", Simona Rabinovici-Cohen, Tal Tlusty, Ami Abutbul, Kari Antila, Xosé Fernandez, Beatriz Grandal Rejo, Efrat Hexter, Oliver Hijano-Cubelos, Abed Khateeb, Juha Pajula, Shaked Perek, published in Proceedings of SPIE Medical Imaging, Vol. 11318, 2020, https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11318/113181B/Radiomics-for-predicting-response-to-neoadjuvant-chemotherapy-treatment-in-breast/10.1117/12.2551374.full
- Paper: "Multimodal Prediction of Breast Cancer Relapse Prior to Neoadjuvant Chemotherapy Treatment", Simona Rabinovici-Cohen, Ami Abutbul, Xosé Fernandez, Oliver Hijano-Cubelos, Shaked Perek, Tal Tlusty, published in Proceedings of International Workshop on PRedictive Intelligence In MEdicine (PRIME-MICCAI), Vol 12329, pages 188-199, 2020, https://link.springer.com/chapter/10.1007/978-3-030-59354-4_18
- Poster: "BigMedilytics - Breast Cancer Pilot", Juha Pajula, Kari Antila, Harri Polonen, Simona Rabinovici-Cohen, Oliver Hijano-Cubelos and Mark van Gils, OpenTech AI 2019 Workshop, https://developer.ibm.com/opentech/2019/03/25/helsinki-may-2019-opentech-ai-workshop/

4.2 Open Source & Resources (refer to ELG)

- We use open source, but didn't create open source from our own technology.
- We created the ISO/IEC 23681:2019 standard: "Self-contained Information Retention Format (SIRF)", https://www.iso.org/standard/76648.html

4.3 Demos

- Demo: "Experiments Viewer for Multimodal Medical Analytics", Simona Rabinovici-Cohen, Tal Tlusty, Ami Abutbul, Efrat Hexter, Abed Khateeb, Shaked Perek, in SPIE Medical Imaging Live Demonstrations Workshop, 2020
- Presentation: "SNIA Long Term Retention for AI Applications", in SDC EMEA 2020, https://www.snia.org/events/sdcemea/agenda