Chapter in book ISBN 978-1-63828-236-5, E-ISBN 978-1-63828-237-2, in press, 23rd January 2024

Pilot 8: Breast Cancer Pilot

Simona Rabinovici-Cohen
IBM Research - Israel, University of Haifa Campus, Mount Carmel, Haifa, Israel

Abstract

Breast cancer remains one of the most widespread and deadly cancers among women today. Early detection and effective treatment of this disease can improve the prognosis significantly. Women with locally advanced breast cancer are generally given neoadjuvant chemotherapy (NAC), in which chemotherapy and optionally targeted treatment is administered prior to surgery. However, clinicians have difficulty estimating the outcomes of NAC in advance. Indeed, experience has shown very different outcomes even across patients with very similar prognostic factors. We report here a study into NAC outcomes prediction using artificial intelligence (AI) on multimodal data of different types, including imaging and other clinical data. Using a cohort of 1738 anonymised patients with breast cancer who received NAC between 2012 and 2018, and a model-to-data (remote visitation) approach, a retrospective study evaluated the prediction of several outcomes of the NAC treatment which were deemed important by the clinicians. Further, we tested our methods in an external competition, BMMR2 Challenge, to validate its generalizability. We won the second place in this competition, with a very small margin from being first and a standout from the other challenge entries. We found that a combination of AI-based techniques and multimodal diagnostic data is therefore a strong contender for improving clinical treatment choices for women with breast cancer.

Keywords: breast cancer; neoadjuvant chemotherapy (NAC); multimodality; medical imaging; magnetic resonance imaging (MRI); artificial intelligence (AI); machine learning (ML); deep learning (DL); image processing; radiomics

0.1.1 Introduction

Breast cancer remains one of the most widespread and deadly cancers among women today [1]. Neoadjuvant chemotherapy (NAC), in which chemotherapy and optionally targeted therapy are administered prior to surgical therapy, is one of the approaches used to treat locally advanced breast cancer. Today, the clinical parameters used to select the NAC option are based on breast cancer subtype, tumour size, disease grade, number of malignant nodes, age, and tumour growth, amongst others [2]. Imaging is being used to evaluate the position of the tumour and its size, but not to predict the outcome of the treatment.

Predicting the outcomes of NAC is an important clinical question. If this future outcome could be predicted based on data available prior to the initiation of NAC treatment, it could impact the treatment selection. However, clinicians have difficulty estimating the outcomes of this treatment prior to its start. In fact, some matching patients have similar prognostic parameters, yet one patient experiences a positive outcome while the other encounters a negative one. Clinicians' treatment selection and decision making could be assisted and

empowered by artificial intelligence (AI) models that could accurately predict NAC outcomes. These AI models are an important enabler of precision medicine.

The breast cancer pilot in BigMedilytics EU project No. 780495 aimed to improve NAC outcomes prediction using artificial intelligence (AI) on multimodal data of different types. We used deep learning (DL) and image processing models for medical imaging data, classical machine learning (ML) models for clinical data, and ensembles of the individual clinical and imaging models. The pilot was a collaboration among Institut Curie in France, VTT in Finland and IBM Research in Israel which also led the pilot. Institut Curie provided the anonymized dataset and clinical expertise, while VTT developed the image processing models, and IBM developed the AI-based multimodal imaging, clinical and ensemble models.

0.1.2 Study design

We created a cohort of 1738 anonymised patients that included women with breast cancer who have received NAC between 2012 and 2018. To comply with regulations as GDPR and French laws, the anonymised dataset was made available to the processing collaborators, through a controlled-access connection to access a local server provided by Institut Curie. We used a model-to-data paradigm where all the data remained at Institut Curie infrastructure. All computations were implemented on a strong GPU enabled server that resided in Institut Curie, and various pipelines of analytics models were transferred to the server and executed there.

In the study, we explored the prediction of several outcomes of the NAC treatment which were deemed important by the clinicians. The NAC treatment includes six months of chemotherapy and optionally targeted therapy, followed by a surgical therapy. The figure below depicts the significant NAC outcomes that we tried to predict prior to the chemotherapy start. It includes (1) pathologic complete response (pCR) at time of surgery, that is achieved for about 30% of the patients, (2) return of cancer in the same location (relapse), which occurs for about 10% of the patients, (3) return of cancer in a distant location (metastasis), which occurs for about 7-8% of the patients, (4) cancer recurrence (relapse or metastasis) within five years since disease diagnosis, which occurs for about 16% of the patients. Note that the first outcome, pCR, is a positive outcome while the other three are negative ones and may suggest treatment reassessment.

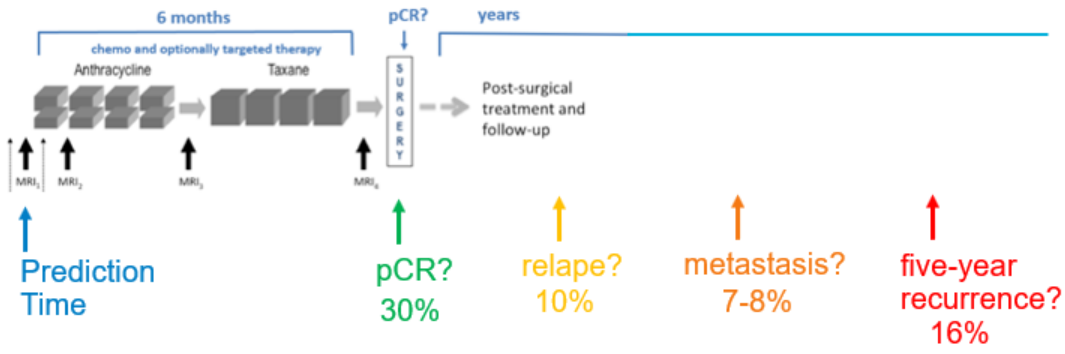

Figure 1: Significant outcomes in neoadjuvant chemotherapy treatment. Accordingly, the pilot explored four prediction tasks: pCR, relapse, metastasis, and five-year recurrence.

0.1.3 Methods

We worked with a real-world retrospective dataset of patients, composed entirely of women diagnosed with breast cancer who had received NAC. The data of each patient included clinical information such as height, weight, age, histological type of the tumour, progesterone status, and many more features. We consider all these data as a single clinical modality. Some of the patients also had in their record medical imaging acquired prior to NAC initiation, which are considered a second modality. Our dataset had labels for the four treatment outcomes that we tried to predict: pCR, relapse, metastasis, and five-year recurrence. However, not all patients had all four labels, and there were some missing values. Given that we have different sizes of datasets for the different modalities and different tasks, our overall multimodal method for the four prediction tasks was as follows (see Figure 2 below). We divided our model into two branches. One branch was trained using clinical data and images, while the other branch was trained using only clinical data. We then combined the two branches into one final ensemble model. To evaluate the models, we performed cross-validation and computed the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) with confidence interval (CI), as well as measured sensitivity, specificity, and other metrics.

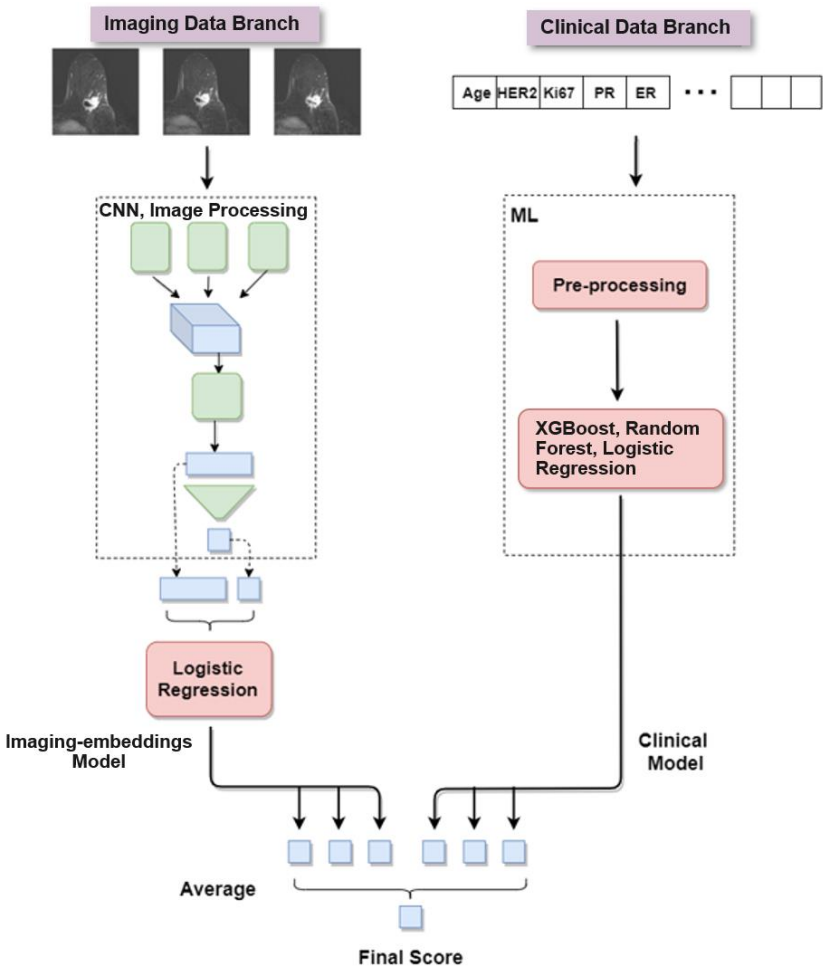

Figure 2: Overall multimodal method for the four prediction tasks. The left branch is trained using clinical data and images. The right branch is trained using clinical data only. The two branches are combined into one final ensemble model.

Clinical Model

The clinical model was similar for all four prediction tasks. We split the cohort with clinical information into five folds with equally distributed positive and negative samples among folds. To select the best classifier for our task, we pre-processed and modelled the data with three known machine learning algorithms: random forest, logistic regression, and XGBoost. The pre-processing included a scaler that scaled all features to the [0, 1] range and an imputation process to replace missing values with the mean value. Since our data were highly unbalanced, we used sample weighting that is inversely proportional to the class frequencies in the input data for the random forest and logistic regression classifiers. For XGBoost, we used positive scaling that is proportional to the ratio between negative and positive samples.

Imaging Model

Interestingly, there wasn't one imaging algorithm that fits all four prediction tasks, but instead each task required a different approach and algorithm to achieve improved performance. For predicting pCR, we used mammography imaging (MG). We detected the tumour using a pretrained model, and then extracted radiomics features from the tumour area. For predicting relapse, we used dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI). We annotated the most important subtraction volume and the significant slice in which the tumour was the largest, and then applied a DL method to train the imaging data. For predicting metastasis, we used both DCE-MRI and MG. Using a DL method, we automatically estimated tumour depth of invasion from the 3D MRI, and using the clinical reports, we extracted tumour size measured in 2D MG. For predicting five-year recurrence, we used multiparametric MRI including DCE, Dixon, and apparent diffusion coefficient (ADC) volumes of MRI. We used both DL and image processing techniques to get improved results, and we also interpreted the features' contribution.

High sensitivity is important in our problem setting since this is the operation point used in clinical practice. It is also important to achieve good specificity at these high sensitivity operation points. Adding the medical imaging to our AI models enabled us to improve the specificity at high sensitivity operation points. This signifies the importance of using medical imaging in the AI models that are going to be deployed in clinical practice.

Ensemble Model

The ensemble model was similar for all the four prediction tasks. It received six scores per patient: three scores based on clinical data and three scores based on the imaging data. To improve generalizability, we created multiple variations of each model, in which a different variation started its training with a different seed. Thus, the three scores for clinical data are produced from three clinical models' variations that differ in their training seed initialization, and the three scores for MRI data are produced from three MRI models' variations. We then examined several strategies for combining and 'ensembling' the models. We first tried the stacking classifier in which we trained a meta model on top of the six models' scores. We also tried several voting strategies, in which some of them consider the threshold of individual models. However, we found that the most effective strategy used the mean value of all available scores per patient, so this became the selected option.

0.1.4 Results

In this section, we briefly describe the results in each one of the four prediction tasks, as well as in the BMMR2 external competition. The results of each prediction task are also associated with a publication that we reference for more detailed description.

Predict pCR

A patient achieves pathologic complete response (pCR) if in the surgery following chemotherapy, an invasive residual tumour in the breast, and invasive disease in the axillary nodes are both absent. Achieving pCR after NAC is correlated with improved disease-free state and overall survival compared with those experiencing a partial or no response to NAC. We developed several models for the task of predicting pCR post NAC treatment and published some of our results in [3]. We created a clinical model, an MG model that is based on mammography images, and an ensemble model that combines the clinical and imaging models.

In our dataset, 528 patients had MG scans, and we found that with this limited amount of data, we could not create a robust deep learning model that directly predicts pCR. We selected instead a different approach. We utilized a deep learning model that was pretrained on IBM proprietary data, which consists of thousands of annotated mammograms to classify the existence of a tumour. That model extracted a heatmap in Curie MG images which represents the tumour detection. We then extracted radiomics texture features from the tumour area and the peritumoral margin of the tumour. The final step in the imaging model was to apply a Random Forest classifier on the extracted radiomics features from the MG imaging. The figure below shows the output of the detection on an MG image and the tumour margins we used for radiomics feature extraction.

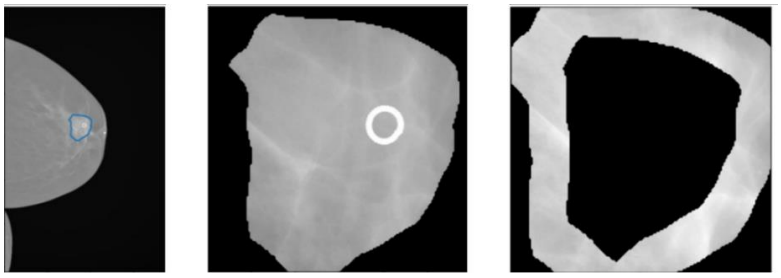

Figure 3: Network output predictions of tumour detection. Left: MG image from Curie dataset with a detected contour around tumour area. Middle: Tumour patch extracted from detected area. Right: Tumour margins extracted.

The final ensemble model combined six models: three models based on clinical data and three models based on features extracted from the MG images. It achieved 0.708 AUC and sensitivity 0.954 while maintaining good specificity of 0.222.

Predict Relapse

A patient encounters relapse if after treatment the breast cancer reoccurs in the same breast. We created multimodal AI models that analyse MRI and clinical data. For the MRI model, expert radiologists annotated the most important subtraction volume in which the tumour appeared to be the brightest in terms of relative illumination. In the selected volume, they

also annotated the significant slice in which the tumour was the largest. Our MRI model included a convolutional neural network (CNN) that 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.

Figure 4 below shows the cross-validation and the holdout test ROC curves for the various models. They exhibit similar trends. In both, the MRI model shows promise in predicting relapse after NAC treatment with good specificity for above 0.95 sensitivity. The clinical model shows the ability to predict relapse with higher specificity around the 0.5 sensitivity but lower specificity around the 0.95 sensitivity. The ensemble of MRI and clinical leveraged both modalities and improved the AUC and specificity at various operation points achieving AUC of 0.735 and specificity of 0.44 on the holdout dataset. The full description of the models and the results were published in [4].

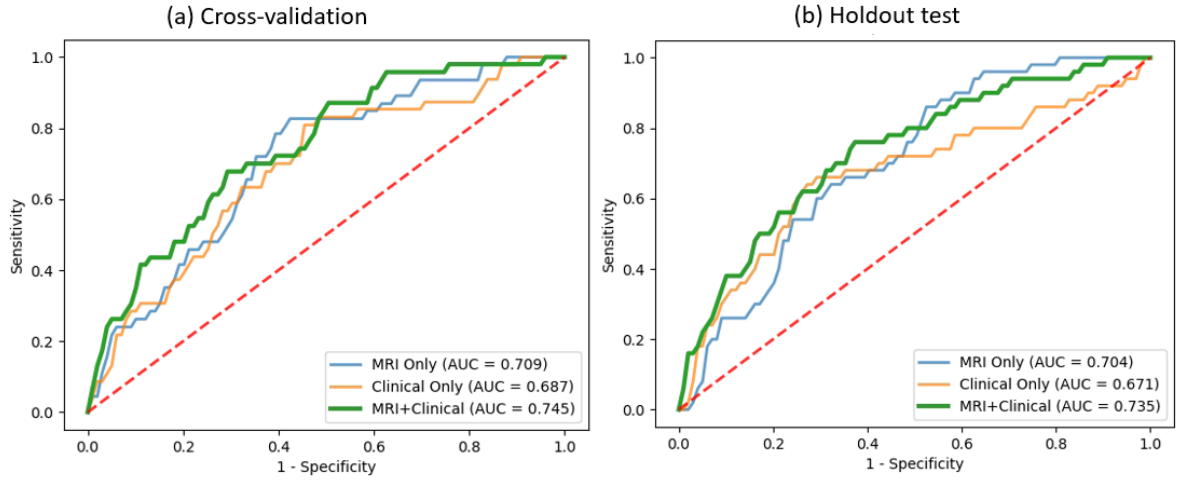

Figure 4: Cross-validation and holdout ROC curves. (a) Cross-validation evaluation with MRI+Clinical ensemble model mean AUC of 0.745 (b) Holdout evaluation with MRI+Clinical ensemble model mean AUC of 0.735.

Predict Metastasis

A patient encounters metastasis if after treatment the breast cancer reoccurs in other areas in the body. We explored the use of tumour size explainable features computed from multimodal imaging and combining it with clinical data to predict the risk of post treatment metastasis. Tumour depth of invasion was automatically estimated from 3D MRI subtraction volumes using a deep learning method that classifies the range of slices in which the tumour is seen and the significant slice. Tumour size as seen in 2D mammography and in clinical examination were extracted from reports. As the patients that have MRI and the patients that have MG only partially overlap, we created a separate model per modality and then ensembled the three models. The ensemble model that combined MRI, MG and clinical data significantly improves the per-modality model as shown in the table below.

Table 1: 5-fold cross-validation evaluation of the per-modality models as well as the ensemble model to predict metastasis.

	Cohort Size	AUC	Spec at Sens=0.95
MRI	551	0.643	0.252
MG	498	0.610	0.166
Clinical	1738	0.649	0.271
Ensemble (MRI Cohort)	**551**	**0.745**	**0.440**

Our method to estimate the tumour depth from MRI scans is fully automatic, and thus more relevant for clinical practice. Moreover, an important aspect of tumour sizes is that these are explainable features, and thus a model based on these predictive features is more likely to be adopted in clinical practice. The full description of the models and the results were published in [5].

Predict Five-Year Recurrence

We say that a patient encounters five-year recurrence if after treatment the breast cancer recurs either locally in the breast (relapse) or distant in other areas of the body (metastasis) within five years from diagnosis. We explored the use of clinical and multiparametric magnetic resonance imaging (mpMRI) to predict the risk of post-treatment recurrence within five years. The mpMRI model uses multiple volumes of the same study and consists of two components. The first component is based on deep learning features extracted from DCE subtraction volumes as done for predicting relapse. The second component is based on traditional image processing methods on Dixon and ADC volumes to generate morphological and texture volumetric features. The final ensemble model that combined clinical and mpMRI models achieved in cross-validation 0.750 [0.698, 0.796] AUC and 0.466 specificity at 0.95 sensitivity operation point, while in the holdout test it achieved 0.734 [0.680, 0.781] AUC and 0.413 specificity.

We also use interpretability methods to explain the model and identify important clinical features for predicting recurrence that when combined can serve as novel candidate composite biomarkers. The figure below provides an explanation of the clinical model via the Shapley Additive Explanations (SHAP) algorithm. SHAP considers all possible combinations of features with and without a specific feature to evaluate its contribution to the prediction. It reveals each feature's importance and demonstrates how each feature of each patient affects the predictive model's results. The figure depicts the top 10 clinical features in descending order that had the most influence on the five-year recurrence prediction. A positive SHAP value means positive impact on the prediction, while a negative value leads the model to predict 'recurrence-free'. The colour of each point

represents the values that each feature can take, including red for high values, blue for low values, and purple for values that are close to the average value.

The categorial clinical features in the data can take the following values: HER2: 0-HER2 negative, 1-HER2 positive; Histological type: 1-NST, 2-lobular, 3-medullary, 4-other; Progesterone status: 0-progesterone negative, 1-progesterone positive; Mitotic index: number of mitoses; and Cancer subtype: 1-TNBC, 2-LuminalA, 3-LuminalB, 4-HER2+.

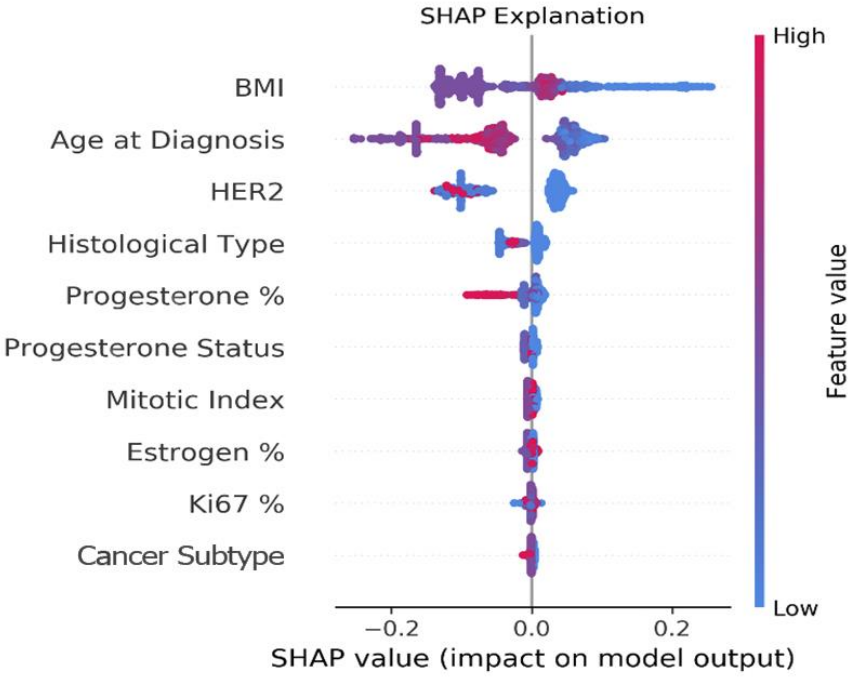

Figure 5: Clinical feature contribution. A summary plot of the SHAP values of the top features in the clinical model. Each point represents a single patient.

Interestingly, Body Mass Index (BMI) and Age at Diagnosis are ranked highest in terms of association with the outcome. In particular, lower values of BMI as well as younger age at the time of diagnosis tend to be associated with a higher risk of five-year recurrence. The full description of the models and the results including interpretation and sub-group analysis were published in [6].

BMMR2 Challenge

We used technologies developed in the breast cancer pilot to validate them in an international external challenge, Breast Multiparametric MRI for prediction of NAC Response (BMMR2) [7], organized by the Breast Imaging Research Program of UCSF[1]. The competition was aiming at predicting pCR based on retrospective analysis of a multicentre clinical trial of cancer patients who completed neoadjuvant chemotherapy prior to surgery. In the competition, IBM were placed second (AUC 0.8380) only marginally lower than the value from Penn Medicine (AUC 0.8397). The open-source technology that the team shared, called FuseMedML [8], a PyTorch-based deep learning framework for medical data, played

[1] https://www.ucsf.edu/

a significant role in the team's ability to quickly experiment with multiple different models and variations and select the best performing one.

0.1.5 Conclusion and Discussion

In this pilot, we explored the prediction of future outcomes in women with locally advanced breast cancer who are treated with NAC. We introduced multimodal prediction models that are based on clinical data and medical imaging taken prior to NAC treatment. Our results demonstrated the ability to predict outcomes prior to NAC treatment initiation using each modality alone. However, a multimodal, ensemble model offers better results. We used deep learning and image processing algorithms to analyse our imaging data and classical machine learning algorithms to analyse 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.

Imaging analysis is generally done via deep neural networks with millions of parameters that need to be learned. Training such a network generally requires thousands of image data and some annotations on the images relating to thousands of patients. However, the medical imaging data available for analytics is scarce, confidential and access to it is protected and limited. Moreover, in medical imaging, the annotations require medical expertise, are expensive, time consuming and inconsistent. Finally, in the medical domain, there is a diversity of populations, genetic variations and environmental differences that may have an impact on the features exhibited in the imaging, and this effect is not quite understood yet. As a result of all these challenges with analysing medical imaging, the creation of robust AI models needs to consider new advanced approaches. Pre-trained models and transfer learning that reuse models trained on external datasets, and federated learning that trains simultaneously on multiple protected datasets can be beneficial approaches to increase the usable dataset and address the medical imaging AI challenges.

In medical imaging AI, multiple modalities are needed as different features are exposed in different modalities. For example, breast density shows up on mammography images but not on ultrasound images, breast calcifications show up on mammography but typically not via ultrasound and never show up on MRI. Thus, multimodal AI models have the potential to provide better performance, and we need to create frameworks and tools for multimodal analysis, such as the FuseMedML open source [8], to ease the research of multimodal analytics.

Medical data is complex. It includes different types such as structured data, text data, genomic data, imaging of different modalities (Xray, MRI, Ultrasound, CT, pathology, and more). Understanding all these modalities and different types of data is complex and requires special expertise. Even within the same modality, different medical centres create different data. For example, MRI has no standardized protocol for scan acquisition and high variance of image resolution, voxel size, and image contrast dynamics. This diversity of modalities increases the data complexity and require special pre-processing and selecting different methods per modality.

AI models that may affect the treatment selection, have direct impact on the patient health, and must be first validated and tested in clinical trial, and then approved by the regulatory authorities such as the FDA in the US and the EMA in Europe. This makes the clinical validation long and difficult, and thus only few validation cycles are possible. Additionally, to increase the acceptance of the AI models, the stakeholders need the ability to interpret the models and understand their reasoning. In our pilot, we provided explanations of our models via the SHAP algorithm as well as via other methods as described. SHAP considers all possible combinations of features with and without that specific feature to evaluate its contribution to the prediction.

Some of our methods were further reused in a following EU Horizon 2020 project, named CAncer PAtients Better Life Experience (CAPABLE). In CAPABLE, we developed AI models to predict 3- and 5-year overall survival rates for patients with metastatic renal cell carcinoma (mRCC). The proposed predictive model, which was constructed as an ensemble of three individual predictive models, outperformed all well-known mRCC prognostic models to which it was compared [9].

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLO-BOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J. Clin. **2021**, 71, 209–249.
2. Fitzgibbons, P.L.; Page, D.L.; Weaver, D.; Thor, A.D.; Craig Allred, D.; Clark, G.M.; Ruby, S.G.; O'Malley, F.; Simpson, J.F.; Connolly, J.L.; et al. Prognostic factors in breast cancer. Arch. Pathol. Lab. Med. **2000**, 124, 966–978.
3. Rabinovici-Cohen, S.; Tlusty, T.; Abutbul, A.; Antila, K.; Fernandez, X.; Grandal Rejo, B.; Hexter, E.; Hijano Cubelos, O.; Khateeb, A.; Pajula, J.; Perek, S. Radiomics for predicting response to neoadjuvant chemotherapy treatment in breast cancer. Proc. of SPIE 11318, Medical Imaging **2020**: Imaging Informatics for Healthcare, Research, and Applications, Vol. 11318.
4. Rabinovici-Cohen, S.; Abutbul, A.; Fernandez, X.; Hijano Cubelos, O.; Perek, S.; Tlusty, T. Multimodal Prediction of Breast Cancer Relapse Prior to Neoadjuvant Chemotherapy Treatment. Proc. of International Workshop on Predictive Intelligence in Medicine (PRIME-MICCAI) **2020**, 188-199.
5. Rabinovici-Cohen, S.; Tlusty, T.; Fernández, X.M.; Grandal Rejo, B. Early prediction of metastasis in women with locally advanced breast cancer. Proc. of SPIE 12033, Medical Imaging **2022**: Computer-Aided Diagnosis, Vol. 12033.
6. Rabinovici-Cohen, S.; Fernández, X.M.; Grandal Rejo, B.; Hexter, E.; Hijano Cubelos, O.; Pajula, J.; Pölönen.; Reyal, F.; Rosen-Zvi, M. Multimodal Prediction of Five-Year Breast Cancer Recurrence in Women Who Receive Neoadjuvant Chemotherapy. Journal of Cancers **2022**, 14(16):3848.
7. Breast Multiparametric MRI for prediction of NAC Response Challenge (BMMR2 Challenge) - The Cancer Imaging Archive (TCIA) Available online: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=89096426 (accessed on 11 March 2023).

8. Golts, A.; Raboh, M.; Shoshan, Y.; Polaczek, S.; Rabinovici-Cohen, S.; Hexter, E. FuseMedML: a framework for accelerated discovery in machine learning based biomedicine. Journal of Open Source Software **2023**, 8 (81).
9. Barkan, E.; Porta, C.; Rabinovici-Cohen, S.; Tibollo, V.; Quaglini, S.; Rizzo, M. Artificial intelligence-based prediction of overall survival in metastatic renal cell carcinoma. Journal of Frontiers in Oncology **2023**, 16;13:1021684.