Proceedings of Machine Learning Research 204:1–15, 2023 Conformal and Probabilistic Prediction with Applications

An Uncertainty-Aware Sequential Approach for Predicting Response to Neoadjuvant Therapy in Breast Cancer

Alberto García-Galindo AGARCIAGALI@UNAV.ES
Marcos López-De-Castro MLOPEZDECAS@UNAV.ES
Rubén Armañanzas RARMANANZAS@UNAV.ES
Institute of Data Science and Artificial Intelligence (DATAI), Universidad de Navarra, Ismael Sánchez Bella Building, Campus Universitario, 31009 Pamplona, Spain
TECNUN School of Engineering, Universidad de Navarra, Donostia-San Sebastián, Spain

Editor: Harris Papadopoulos, Khuong An Nguyen, Henrik Boström and Lars Carlsson

Abstract

Neoadjuvant therapy (NAT) is considered the gold standard preoperative treatment for reducing tumor charge in breast cancer. However, the tumor's pathological response highly depends on patient conditions and clinical factors. There is a dire need to develop modeling tools to predict a patient response to NAT and thus improve personalized medical care plans. Recent studies have shown promising results of machine learning (ML) methodologies in breast cancer prognosis through the combination of several modalities, including imaging and molecular features derived from biopsy analyses. We here present a ML model to predict response to NAT through two sequential prediction stages. First, a pre-treatment dynamic contrast-enhanced magnetic resonance imaging model is trained, followed by a second model with molecular biomarkers-enriched data. We propose the integration of the Conformal Prediction (CP) framework in the first non-invasive model to identify patients whose predicted responses show large uncertainty and refer them to the second model that includes data from invasive tests. The major advantage of this procedure is in the reduction of unnecessary biopsies. Different alternatives for the standard ML algorithms and the CP functions are explored on a publicly available clinical dataset. Results clearly show the potential of our uncertainty-aware clinical predictive tool in such real scenarios.

Keywords: Breast cancer · Neoadjuvant therapy · Pathological response · DCE-MRI radiomics · Molecular biomarkers · Uncertainty · Conformal Prediction

1. Introduction

Breast carcinoma is considered the most prevalent type of invasive cancer among women worldwide (Łukasiewicz et al., 2021), becoming a major public health issue. Despite significant progress in recent decades to reduce mortality rates, there is a critical need to develop better prognostic tools and therapeutic approaches that can improve patient outcomes. In this sense, neoadjuvant therapy (NAT) has become one of the gold standard pre-operative treatments to reduce tumor burden and improve the patient's chances of breast conserving surgery rather than mastectomy (Selli and Sims, 2019). There are several NAT modalities, including (a) chemotherapy, (b) radiation therapy, and, (c) endocrine therapy, and their administration depends on the patient's characteristics and clinical factors. In the best-case scenario, a patient treated with NAT achieves pathological complete response, i.e., absence

of any residual invasive disease (Cortazar and Geyer, 2015). Such optimal response to NAT is far from guaranteed. In fact, several studies (see for example Spring et al. (2020); Romeo et al. (2021)) report that only about 20 - 30 % of breast cancer patients achieve complete response, with success rates varying depending on tumor biology. Consequently, those patients experiencing ineffective NAT incur in toxicity and side effects without reaching the desired clinical benefits. In this context, the development of predictive tools to early assess whether a patient will achieve pathological complete response becomes key.

Several biomarkers have been analyzed to predict response to NAT in breast cancer patients through machine learning (ML) algorithms. These include clinical and molecular predictors (Goorts et al., 2017), as well as pre-treatment dynamic contrast-enhanced medical resonance imaging (DCE-MRI) features (Mani et al., 2013; Massafra et al., 2022). Each of these methods has its own strengths and limitations. One the one hand, pre-treatment DCR-MRI can provide a fast, early and non-invasive assessment of tumor response with no additional cost to those patients where MRI is part of their preoperative test (Cain et al., 2019). On the other hand, molecular-level predictors identified through biopsy can provide better understanding of biological processes with the drawback of being invasive and incurring risk of patient infection.

We argue that non-invasive imaging protocols should be preferred in favor of quality patient care. Only in those cases where the imaging features lack meaningful predictive power, should a biopsy be performed. However, standard ML and data-driven models provide no reliable levels of individual uncertainty and are unable to determine the confidence of a particular prediction. To overcome this limitation, the Conformal Prediction (CP) (Vovk et al., 2022; Balasubramanian et al., 2014) framework offers statistical procedures to generate set predictions that are guaranteed to contain the ground truth with a user specified error rate and minimal assumptions. Set predictions with high cardinality point to samples that are difficult to predict, while set predictions with unique labels are associated with the most confident cases. The CP framework has been applied across many medical disciplines, including breast cancer diagnosis (Lambrou et al., 2009) and survivability (Alnemer et al., 2016), but CP models to predict NAT response remain uncovered.

1.1. Contribution

In this work, we propose the integration of CP within the design of a multimodal ML model for predicting three possible severity responses to NAT in breast cancer patients. Specifically, our proposal lies on a sequential ensemble defined by two chained predictors: a first-stage model crafted on the basis of DCE-MRI data and a second-stage model augmented with molecular features. The key idea is to take advantage of the uncertainty quantification property of CP to (1) predict responses to NAT through the radiomics model only on those cases for whom the model is sufficiently confident, and, (2) ask for a "second opinion" for the most ambiguous patients and compute a prediction using the augmented model. Hence, our ML model allows to avoid unnecessary invasive tests and improve overall patient care management.

We present an application of the proposed sequential model using a publicly available dataset that suits the setting of our clinical problem.

2. Materials and methods

In this section, we introduce the clinical dataset we employ to train and validate our proposed ML solution. We specify the cohort selection and define the predictive task in medical terms, including the clinical target. Then, we briefly review the inductive version of the CP framework within the classification paradigm as the underlying method of our proposal. Finally, we present the detailed aspects of our model, including training and new patient assessment workflows.

2.1. Patient population and problem definition

The Duke Breast Cancer MRI dataset (Saha et al., 2018), available through the Cancer Imaging Archive (TCIA) (Clark et al., 2013), provides a fully annotated and anonymazed collection of 922 invasive breast cancer patients admitted at Duke University Hospital between January 1st, 2000 and March 23rd, 2014.

Out of this cohort, we identified a subset of 312 cases treated with NAT. To evaluate the effectiveness of NAT, pathological reports from the first surgical intervention were obtained. Such reports allowed to further characterize NAT tumor responses according to the *ypTNM* re-staging classification. In clinical practice, the *ypTNM* classification is a widely recognized and standard system for measuring cancer re-staging after adjuvant therapies, based on the size of the primary tumor (T), the presence and extent of lymph node involvement (N), and the presence of distant metastasis (M) (Cserni et al., 2018). We identified 240 patients treated with NAT with complete *ypTNM* classification status available. From this classification, each patient was matched to a meaningful global cancer stage, ranging from IA (only a small local tumor mass remains present) to IV (tumor has spread to other distant organs, i.e., metastasis).

We considered three possible severity responses to NAT therapy based on representative sample sizes and clinical relevance: (a) pathological complete response, (b) early stage, and, (c) locally advanced or metastatic stage. These responses were defined according to Table 1 criteria.

Table 1: Clinical definitions and sample sizes for severity response to NAT.

Response	**Cancer re-stage**	**Sample size**
Pathological complete response	$\varnothing$	71 (29.6 %)
Early stage	IA or IIA	104 (44.3 %)
Locally advanced or metastasis stage	From IIB to IV	65 (27.1 %)

Patient's age, menopausal status and different information modalities to address the response to NAT predictive task were also identified, including:

- **Imaging sequences and feature extraction** For each patient in the final cohort, pretreatment axial breast DCE-MRI assessments were available from 1.5T or 3T scanners. Several sequences were acquired: (a) T1-weighted fat-satured pre-contrast sequence, (b) T1-weighted fat-satured sequence after contrast agent administration, and, (c) T1-weighted non-fat satured sequence. Once the sequences were recorded,

breast tumors were manually segmented by radiology experts using three-dimensional boxes. Then, a tumor mask was obtained applying a fuzzy C-means algorithm inside the annotations boxes. Potentially meaningful radiomic features were extracted using an inhouse software from Duke University Hospital. Specifically, a comprehensive list of 539 potential image biomarkers describing texture and time-dependent tumor and fibroglandular tissue characteristics are available for each patient. A detailed explanation of each radiomic feature is available on Saha et al. (2017).

- **Pathological information** Several features derived from tumor genomics and molecular profiling were determined through pathology assessments and immunohistochemical analyses from the biopsy. We here considered three main receptors commonly tested in breast cancer whose status provides key information about cell division and tumor growth: estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor (HER2). We further include the surrogate molecular subtype from such receptor status: Luminal A, Luminal B, HER2+, and triple-negative. Additionally, we considered several tumor biological features, including:

  - Histological type, describing tumor tissue and cell morphology.
  - Histological grade according to the Nottingham Grade System, which is based on tubule formation, nuclear pleomorphism, and mitosis rate.
  - Initial cancer staging based on the TNM classification system.

  A total set of 12 features describing tumor biology were identified for each patient.

2.2. Inductive CP

The CP framework provides statistical procedures to quantify reliable levels of individual uncertainty by means of multivalued prediction regions. Let's assume a set formed by a sequence of training samples $\{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$. Given a new test object x_{m+1}, a conformal predictor compute a prediction interval $\Gamma_\epsilon(x_{m+1})$ that will contain the true y_{m+1} with user specified expected error rate $\epsilon \in [0, 1]$.

The development of a conformal predictor rests on the definition of a non-conformity measure $\mathcal{S} : \mathcal{Z} \times \mathcal{Z}^{(k)}$, which is employed to quantify the degree of strangeness of a sample z with respect to a given collection $\{z_1, \ldots, z_k\}$. Although the non-conformity measure can be by definition any measurable function, it is usually shaped from the output of a ML model. Formally,

$$\mathcal{S}(z, \{z_1, \ldots, z_k\}) = \Delta(y, h(x)) \tag{1}$$

where $h : \mathcal{X} \to \mathcal{Y}$ is a predictive model learned on $\{z_1, \ldots, z_k\}$ and $\Delta : \mathcal{Y} \times \mathcal{Y}$ is a function of dissimilarity between the real target y and the prediction $h(x)$.

CP was originally proposed through a transductive strategy, in which the non-conformity measure needs to be computed for each training sample z_i. In practice, the transductive strategy is computationally demanding since it requires the underlying ML model to be retrained for each test object (Vovk, 2013).

In this work, we focus on inductive CP (ICP) (Papadopoulos et al., 2002), a modified approach to build conformal predictors in an efficient way. Under the ICP scheme, the

training set is split into a proper training set $(z_1, \ldots, z_n)$ and a calibration set $(z_{n+1}, \ldots, z_m)$. The proper training set is used to learn a single predictive model. This predictive model will be used to compute the non-conformity measure on the remaining calibration samples. At this point, we can assign a new test object x_{m+1} with a hypothetical value $\overline{y}_{m+1}$ and check how strange this completion $z_{m+1} = (x_{m+1}, \overline{y}_{m+1})$ is with respect to the calibration samples. Therefore, given the non-conformity scores:

$$\begin{aligned} s_i &= \mathcal{S}(z_i, \{z_1, \ldots, z_n\}) \qquad i = n+1, \ldots, m \\ s_{m+1} &= \mathcal{S}(z_{m+1}, \{z_1, \ldots, z_n\}) \end{aligned} \tag{2}$$

we can compute a p-value for each label as follows:

$$p_{\overline{y}_{m+1}} = \frac{|\{i = 1, \ldots, m+1 : s_i > s_{m+1}\}|}{m+1} \tag{3}$$

For a given error rate $\epsilon \in [0, 1]$, the prediction region for the test object x_{m+1} will contain all the possible target values $\overline{y}_{m+1}$ whose p-value is greater than ϵ:

$$\Gamma_\epsilon(x_{m+1}) = \{\overline{y}_{m+1} \,|\, p_{\overline{y}_{m+1}} > \epsilon\} \tag{4}$$

This conformal predictor can be built on top of any supervised ML model, and it provides validity (i.e., true label coverage) without requiring any further assumption beyond exchangeability on the probability distribution (i.e. the probability measure does not depend on the order of its arguments).

Forced single predictions. Prediction sets are usually the output choice for an inductive conformal predictor. However, for a given test sample x_{m+1}, a forced single prediction can be computed by assigning the target value with the highest p-value (i.e., the less non-conformal target value). This allows to produce point predictions by an inductive conformal predictor comparable to conventional ML models.

$$\Gamma_{\mathsf{forced}}(x_{m+1}) = \underset{\overline{y}_{m+1} \in \mathcal{Y}}{\operatorname{argmax}} \, p_{\overline{y}_{m+1}} \tag{5}$$

The forced prediction does not depend on the desired error rate ϵ, and allows to fairly compare the output of an inductive conformal predictor with the point predictions generated from a standard ML model.

2.3. Uncertainty-Aware Sequential Modeling

For the sake of developing a ML approach to predict tumor response to NAT, we rely on a dataset of sample triplets:

$$(x_1^{\mathsf{MRI}}, x_1^{\mathsf{BIO}}, y_1), \ldots, (x_n^{\mathsf{MRI}}, x_n^{\mathsf{BIO}}, y_n) \qquad x_i^{\mathsf{MRI}} \in \mathcal{X}^{\mathsf{MRI}}, \quad x_i^{\mathsf{BIO}} \in \mathcal{X}^{\mathsf{BIO}}, \quad y_i \in \mathcal{Y} \tag{6}$$

where x_i^{MRI} denotes a collection of radiomic features acquired through MRI sequencing, x_i^{BIO} denotes a collection of cancer profile biological features assessed through biopsy and immunochemistry analyses, and y_i denotes the severity response to NAT.

Our goal is to learn and validate a classifier using this dataset. Under a conventional ML approach, one could initially consider the whole set of features to train an algorithm

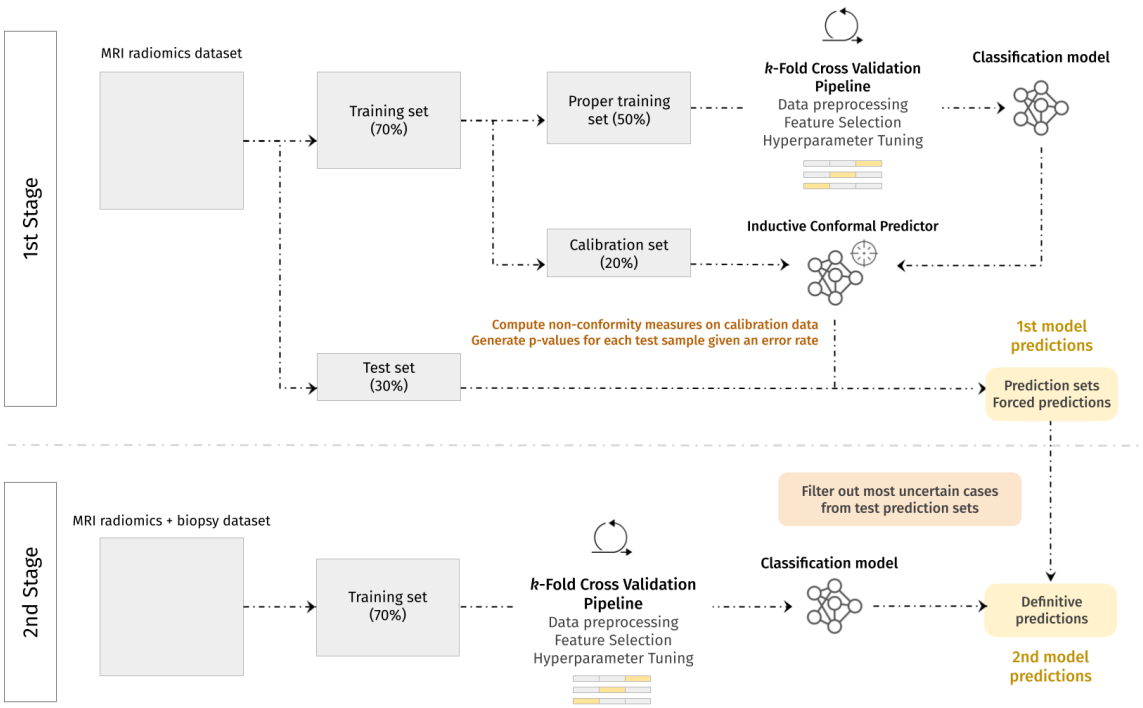

Figure 1: Uncertainty-Aware Sequential Predictive Pipeline data flow

and generate predictions for future patients. However, when it comes to predict response to NAT, each of the considered data modalities is gathered from independent assessments with different associated cost for the patients. Our alternative is a sequential ML model based on the uncertainty quantification property of CP to predict response to NAT in breast cancer patients by efficiently use each data modality.

The pseudocode of our sequential model is presented in Algorithm 1 and visually illustrated in Figure 1. The proposal follows the ensemble of two chained models: a first non-invasive model trained from the MRI data (see 1st Stage from Algorithm 1), and, a second model enriched with clinical information gathered from the biopsy analyses (see 2nd Stage from Algorithm 1). We are interested in predicting a response to NAT using only non-invasive features when the model is confident in the prediction. This is addressed by learning an inductive conformal predictor built on top of such non-invasive features. We use the uncertainty quantification property of CP to generate prediction regions for each patient and identify those whose set size is above what we have called an uncertainty set size threshold. In this way, we can distinguish between confident and uncertain predictions. The confident predictions are kept to be predicted through the conformal predictor, while the uncertain predictions are referred to the second model to produce a final prediction. It can be noticed that both modeling stages must share the same train/test splits to provide a fair end-to-end performance evaluation of the sequential model.

Algorithm 1: Uncertainty-Aware Sequential Model

Input: Patient triplets dataset $\mathcal{D} = \{(x_i^{\mathsf{MRI}}, x_i^{\mathsf{BIO}}, y_i)\}_{i=1}^n \in \mathcal{X}^{\mathsf{MRI}} \times \mathcal{X}^{\mathsf{BIO}} \times \mathcal{Y}$

Supervised learning classification algorithm $\mathcal{A}(\lambda)$.

Hyperparameter configuration grid $\Lambda = \{\lambda^{(1)}, \ldots, \lambda^{(m)}\}$.

Non-conformity measure $\mathcal{S} = \Delta(y, h(x))$.

Error rate $\epsilon \in [0, 1]$.

Uncertainty set size threshold $\tau \in \{0, \ldots, |\mathcal{Y}|\}$

Output: Sequential Predictive Model

1st Stage: Non-invasive MRI model training and calibration

Out of $\mathcal{D}$, select $\mathcal{D}^{\mathsf{MRI}} = \{(x_i^{\mathsf{MRI}}, y_i)\}_{i=1}^n$.

Split $\mathcal{D}^{\mathsf{MRI}}$ into a training set $\mathcal{D}^{\mathsf{MRI}}_{\mathsf{train}}$ and a test set $\mathcal{D}^{\mathsf{MRI}}_{\mathsf{test}}$.

Split $\mathcal{D}^{\mathsf{MRI}}_{\mathsf{train}}$ into a proper training set $\mathcal{D}^{\mathsf{MRI}}_{\mathsf{proper}}$ and a calibration set $\mathcal{D}^{\mathsf{MRI}}_{\mathsf{calib}}$.

Find optimal $\mathcal{A}(\lambda^*), \lambda^* \in \Lambda$ using cross validation over $\mathcal{D}^{\mathsf{MRI}}_{\mathsf{proper}}$.

Learn $h(x^{\mathsf{MRI}})$ from $\mathcal{D}^{\mathsf{MRI}}_{\mathsf{proper}}$ using $\mathcal{A}(\lambda^*)$.

Compute non-conformity measures on calibration set $\{\Delta(y_i, h(x_i^{\mathsf{MRI}}))\}_{i \in \mathcal{D}^{\mathsf{MRI}}_{\mathsf{cal}}}$.

Induce a conformal predictor and infer $\{\Gamma^\epsilon(x_i^{\mathsf{MRI}})\}_{i \in \mathcal{D}^{\mathsf{MRI}}_{\mathsf{test}}}$ according to $\{s_i\}_{i \in \mathcal{D}^{\mathsf{MRI}}_{\mathsf{cal}}}$.

Compute $\Gamma_{\mathsf{forced}}(x_i^{\mathsf{MRI}})$ for every test sample that $|\Gamma^\epsilon(x_i^{\mathsf{MRI}})| \leq \tau$.

2nd Stage: Invasive MRI + biopsy model training

Select $\mathcal{D}_{\mathsf{train}} = \{x_i^{\mathsf{MRI}}, x_i^{\mathsf{BIO}}, y_i\}_{i \in \mathcal{D}^{\mathsf{MRI}}_{\mathsf{train}}}$.

Find optimal $\mathcal{A}(\lambda^*), \lambda^* \in \Lambda$ using cross validation over $\mathcal{D}_{\mathsf{train}}$.

Learn $h(x^{\mathsf{MRI}}, x^{\mathsf{BIO}})$ from $\mathcal{D}^{\mathsf{MRI}}_{\mathsf{train}}$ using $\mathcal{A}(\lambda^*)$.

Compute a final prediction $h(x^{\mathsf{MRI}}, x^{\mathsf{BIO}})$ for every test sample that $|\Gamma^\epsilon(x_i^{\mathsf{MRI}})| > \tau$.

3. Experimental settings and results

In this section, we cover the experimental set up followed to develop and assess the validity of our sequential model. We present two different experiments:

- Experiment 1: Benchmarking assessment. It is worth noting that the motivation of our proposal lies on the assumption that the complete multimodal feature set provide a more meaningful information for predicting response to NAT than considering the MRI radiomic biomarkers alone. Consequently, we conducted an initial experiment to quantify to what extent the biopsy analysis improved the predictive performance with respect to the MRI-only model. These models establish a bottom and a top predictive limits on which our sequential model will range based on the number of filtered patients from the first model to the second model.

- Experiment 2: Uncertainty-aware proposal. Considering the benchmarking results, we trained our sequential model with different patient filtering patterns and compared it

with the baseline models predictive performance, as well as the number of patients referred to the second stage model.

3.1. Data preparation and model training

In both experiments, we followed a common procedure to prepare data, train the classification algorithms, and estimate predictive performance results.

We initially split our dataset into 70 % for training our prediction models, with the remaining 30 % as test set to generate predictions and quantify performance. Both training and test sets followed the same class distribution as the original dataset.

We employed a learning pipeline to preprocess data and train the predictive algorithms. Regarding feature preprocessing, we drop features with a linear correlation coefficient higher than 0.99. We then converted nominal features to one-hot encoded binary features and scaled continuous features to zero-mean and unit variance. We set missing values to the median for continuous attributes and to *unknown* category in the case of nominal features. From the resulting dataset, we then performed an univariate feature selection step by means of a mutual information criteria, in which we obtained a score for each feature; the higher the score, the more important was the feature towards the target variable. The total number of features to be retained by this selection step was considered a pipeline hyperparameter to be tuned. We further considered three different state-of-the-art supervised learning algorithms: *logistic regression* (LR), *random forest* (RF), and *xgboost* (XGB). We chose these algorithms since they represent different modeling paradigms: *logistic regression* is an easily understandable model that assumes linearity of features, *random forest* is based on the foundations of bootstrap aggregation of several decision trees, and *xgboost* is an example of a more complex boosting method. We used standard implementations from Python's `scikit-learn` module (Pedregosa et al., 2011) for the first two algorithms and `xgboost` module (Chen and Guestrin, 2016) to train the gradient-boosted trees.

To find the optimal learning configuration (i.e., k-best ranked features using the mutual information criteria and the classification algorithm hyperparameters) and avoid data leakage, we train the complete learning pipeline using a stratified k-fold cross validated random search over a fixed parameter grid. We did so using a 3-fold cross validation and searching across 50 random pipeline configurations. To account for label imbalance, each algorithm was trained with balanced class-dependent weights.

3.2. Performance evaluation

Once a model was learned, we predicted response to NAT for the test set and report several multiclass metrics to assess the predictive performance of the ML models. We consider both global and class-level metrics to detail the behavior on each of the possible responses to NAT. The complete list of the covered metrics and a brief description are shown in Table 2.

3.3. Conformal predictor design

We now describe the design aspects of the inductive conformal predictor, that is the backbone of our methodological proposal. Note that the following settings only apply to Experiment 2.

Table 2: Standard performance metrics considered in our multiclass predictive task.

Metric	Definition
$\text{Accuracy} = \frac{1}{N}\sum_{i}^{N} \mathbb{1}(y_i = \hat{y}_i)$	Correct classification rate
$\text{Recall}_{\mathsf{CR}} = \frac{1}{N_{\mathsf{CR}}}\sum_{i}^{N_{\mathsf{CR}}} \mathbb{1}(y_i = \mathsf{CR})\mathbb{1}(\hat{y}_i = \mathsf{CR})$	Pathological complete response cases correctly identified
$\text{Recall}_{\mathsf{ES}} = \frac{1}{N_{\mathsf{ES}}}\sum_{i}^{N_{\mathsf{ES}}} \mathbb{1}(y_i = \mathsf{ES})\mathbb{1}(\hat{y}_i = \mathsf{ES})$	Early stage cases correctly identified
$\text{Recall}_{\mathsf{LA}} = \frac{1}{N_{\mathsf{LA}}}\sum_{i}^{N_{\mathsf{LA}}} \mathbb{1}(y_i = \mathsf{LA})\mathbb{1}(\hat{y}_i = \mathsf{LA})$	Locally advanced or metastasic stage cases correctly identified
$\text{F1}_{\mathsf{macro}} = \frac{1}{3}\,\text{F1}_{\mathsf{CR}} + \text{F1}_{\mathsf{ES}} + \text{F1}_{\mathsf{LA}}$	Unweighted F1 between all per-class F1 scores

To build the conformal predictors, we further split the training set into a stratified 50 % proper training set to train the underlying learning pipeline (see Section 3.1) and a stratified 20 % calibration set on which non-conformity scores are computed.

As explained in Section 2.2, the definition of a suitable non-conformity measure is a key element when developing a conformal predictor. In this work, we test two different non-conformity measures: (a) inverse probability error, and, (b) marginal error.

$$\text{IPE}(y_i, h(x_i)) = 1 - h(x_i)_{y_i} \tag{7}$$

$$\text{ME}(y_i, h(x_i)) = 0.5 - \frac{h(x_i)_{y_i} - \max_{y!=y_i} h(x_i)_y}{2} \tag{8}$$

Another important step in our sequential model involves the definition of a sample filtering strategy to refer highly uncertain cases from the non-invasive model to the invasive one. Such filtering depends on two factors: the uncertainty set size threshold $\tau \in \{0, \ldots, |\mathcal{Y}|\}$, and the user specified error rate $\epsilon \in [0, 1]$. Note that, given a fixed τ, testing different error rates lead to different prediction sets and, consequently, to different patient filtering patterns.

We set $\tau = 1$, setting up a conservative threshold for which only single set predictions are kept to be predicted by the first model. In order to produce prediction sets, we test two different error rates: $\epsilon \in \{0.1, 0.2\}$.

To assess the performance of the inductive conformal predictors, we further report several uncertainty metrics to measure both validity and efficiency properties. These metrics are presented in Table 3.

4. Results

4.1. Benchmark performance

Table 4 shows prediction performance metrics from the benchmark experiments each classification algorithm. Table 4(a) includes evaluation results for the MRI models, whereas

Table 3: Uncertainty evaluation metrics considered for measuring inductive conformal predictors performance.

Metric	**Definition**
$\text{Coverage} = \frac{1}{N}\sum_{i}^{N}(y_i \in \Gamma(x_i))$	True label coverage rate
$\text{Empty rate} = \frac{1}{N}\sum_{i}^{N}(\lvert\Gamma(x_i)\rvert = 0)$	Empty prediction set rate
$\text{Multiple rate} = \frac{1}{N}\sum_{i}^{N}(\lvert\Gamma(x_i)\rvert > 1)$	Multiple prediction set rate
$\text{Single rate} = \frac{1}{N}\sum_{i}^{N}(\lvert\Gamma(x_i)\rvert = 1)$	Single prediction set rate
$\text{Single coverage} = \frac{1}{N_{\mathsf{single}}}\sum_{i}^{N_{\mathsf{single}}}(y_i \in \Gamma(x_i))$	True label coverage for single prediction sets
$\text{True single rate} = \frac{1}{N}\sum_{i}^{N}(y_i \in \Gamma(x_i) : \lvert\Gamma(x_i)\rvert = 1)$	Single prediction set covering true label rate

Table 4(b) summarizes classification performance based on the complete feature set. Additionally, Table 4(c) highlights the predictive improvement across metrics achieved by including all the available data.

Table 4: Test set classification metrics for (a) the MRI models, (b) the MRI+BIO models, and (c) the predictive improvement. For stability, we ran the experiment 25 times and averaged metrics over all iterations.

	(a)			(b)			(c)
	$\mathbf{LR_{MRI}}$	$\mathbf{RF_{MRI}}$	$\mathbf{XGB_{MRI}}$	$\mathbf{LR^{BIO}_{MRI}}$	$\mathbf{RF^{BIO}_{MRI}}$	$\mathbf{XGB^{BIO}_{MRI}}$	$\mathbf{\Delta^{BIO}}$
Accuracy	0.424	**0.434**	0.432	0.524	0.532	**0.536**	0.102
$\text{Recall}_{\mathsf{CR}}$	**0.364**	0.217	0.269	**0.488**	0.455	0.360	0.124
$\text{Recall}_{\mathsf{ES}}$	0.434	**0.514**	0.461	0.480	0.539	**0.672**	**0.158**
$\text{Recall}_{\mathsf{LA}}$	0.472	0.538	**0.560**	**0.630**	0.602	0.510	0.070
$\text{F1}_{\mathsf{macro}}$	**0.416**	0.408	0.414	0.523	**0.525**	0.512	0.109

It is noteworthy that the selected features consistently maintained a comparable magnitude across experiments, with an average range of approximately 75 features.

In general terms, the MRI models showed a moderate predictive power regarding overall accuracy and F1 metrics, reaching a maximum of 0.434 and 0.416, respectively. Despite each of the trained algorithms performing quite similar for these metrics, they yielded to different results across recall classes. The LR model performed better for identifying complete responses to NAT (0.364). When predicting early stage cases, the RF model was

the best option (0.514), whereas the XGB model achieved the highest score on the third class (0.560), namely locally advanced cancer patients.

In the case of the MRI+BIO models, the LR model outperformed the other alternatives for identifying minority classes, reaching a 0.488 and a 0.630 recall scores for complete response and locally advanced cases, respectively. The XGB model achieved the best global accuracy (0.536) and a remarkable score of 0.672 for identifying early stage cases, the highest metric overall. Finally, it is worth mentioning that both the RF model and the LR model performed similarly good on F1 metrics (0.525 and 0.523).

Comparing which ML model is best for each scenario, we identified significant gaps for each evaluation metrics. On average, the invasive models improved the MRI classifiers by an 11.2 %. On one hand, the largest difference was observed for the early stage patients, with an improvement of 15.8 %. On the other hand, the prediction of locally advanced cases improved the least (7 %). These results confirm our initial hypothesis that, despite reasonable predictive power, the MRI models underperformed when predicting response to NAT. The inclusion of the biological features provided valuable information to the algorithms, producing improvements on every performance metric, and achieving substantial better results.

4.2. Uncertainty-aware model performance

The uncertainty metrics for the induced conformal predictors are reported in Table 5, whereas predictive performance metrics using the sequential model are shown in Table 6. It is worth mentioning that these performance metrics must be interpreted taking into account the number of patients kept in the first model (i.e., the single rate reported in Table 5).

Table 5: Test set uncertainty metrics for the inductive conformal predictors. Empty set rates were not reported since they were not generated in any of the experiments.

		LR		**RF**		**XGB**	
		$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.1$	$\epsilon = 0.2$
IPE	Coverage	0.911	0.798	0.880	0.768	0.889	0.788
	Multi	0.924	0.813	0.948	0.839	0.958	0.863
	Single	**0.076**	**0.187**	0.052	0.161	0.042	0.137
	Single coverage	**0.873**	**0.772**	0.820	0.733	0.866	0.762
	True single	0.066	0.145	0.043	0.118	0.037	0.105
ME	Coverage	0.822	0.774	0.888	0.782	0.910	0.816
	Multi	0.825	0.663	0.874	0.724	0.879	0.747
	Single	**0.175**	**0.337**	0.126	0.276	0.121	0.253
	Single coverage	0.828	0.735	0.854	0.766	**0.901**	**0.797**
	True single	0.145	0.247	0.108	0.212	0.109	0.201

For the IPE-calibrated models, prediction coverage almost reached the expected theoretical value, either 0.9 or 0.8. Regarding set sizes, the LR conformal predictor achieved the highest rate of patients kept within the first stage for both $\epsilon = 0.1$ (0.076) and $\epsilon = 0.2$

Table 6: Test set predictive performance metrics for the sequential model. Note that this metrics are computed from both 1st stage forced predictions (confident cases) and 2nd stage point predictions (uncertain cases).

		LR		**RF**		**XGB**	
		$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.1$	$\epsilon = 0.2$
IPE	Accuracy	0.519	0.507	0.527	0.518	0.528	0.525
	$\text{Recall}_{\text{CR}}$	0.463	0.438	0.371	0.345	0.314	0.316
	$\text{Recall}_{\text{ES}}$	0.506	0.500	0.604	0.598	0.674	0.659
	$\text{Recall}_{\text{LA}}$	0.598	0.588	0.570	0.576	0.526	0.536
	F1_{macro}	0.515	0.501	0.506	0.494	0.498	0.497
ME	Accuracy	0.494	0.489	0.524	0.508	0.515	0.508
	$\text{Recall}_{\text{CR}}$	0.421	0.410	0.366	0.337	0.320	0.318
	$\text{Recall}_{\text{ES}}$	0.505	0.498	0.564	0.552	0.639	0.627
	$\text{Recall}_{\text{LA}}$	0.554	0.558	0.623	0.618	0.528	0.524
	F1_{macro}	0.486	0.479	0.513	0.493	0.491	0.483

(0.187). This result became more meaningful since the LR sequential model also outperformed the RF and the XGB in classifying complete response and locally advanced cases. Note that the XGB sequential model was the best for predicting early stage cases, and, for $\epsilon = 0.1$, this model produce a single rate of 0.042 without sacrificing predictive power for this class.

If we focus on the ME-calibrated models, we observe an overall increment on the single set rates for each algorithm. Consider, for example, the case of the RF. For the ME-calibrated predictor, single set rate with $\epsilon = 0.1$ (0.052) was more than double the single set rate for the IPE-calibrated predictor with $\epsilon = 0.2$ (0.126). The highest single set rate was at 0.337, reached by the LR conformal predictor with $\epsilon = 0.2$. In addition, the XGB conformal predictor produced the highest single rate coverage for both error rates (0.901 for $\epsilon = 0.1$ and 0.797 for $\epsilon = 0.2$).

Three different models achieved reasonably good overall classification performance in our experiments. First, the ME-RF sequential model, that achieved a global F1 score of 0.513 for $\epsilon = 0.1$ while retaining a 12.6 % of the cases within the first stage modeling. Note that this model achieved a comparable overall predictive power with respect to the MRI+BIO best model performance (see Table 4(*b*)). We also highlight the IPE-XGB sequential model with $\epsilon = 0.2$. This model was the best candidate for predicting early stage cases, reaching a 0.659 recall score while keeping a 13.7 % of the patients assessed by the first stage MRI model, with a competitive F1 score of 0.497. Finally, the IPE-LR sequential model with $\epsilon = 0.1$ was the preferable option for identifying complete response cases. This model achieved a recall score for this class of 0.463, the closest one to the results from the MRI+BIO best model. As opposed to the other two sequential models, the IPE-LR only retained a 7.6 % of the patients within the first stage.

5. Conclusions

Early accurate evaluation of how a breast cancer patient will respond to NAT is key for guiding personalized treatments. Machine learning methodologies have shown promise for such problem by integrating data from different clinical assessments with varying associated cost for the patients. In this study, we propose an uncertainty-aware sequential model that makes an efficient use of different data modalities for predicting response to NAT in breast cancer patients. Our proposal is based on two different prediction stages: a first conformal predictor built on top of non-invasive MRI features, and a second conventional predictive model based on biological features gathered from invasive tests. The integration of conformal prediction into the first modeling stage helps identify patients with large uncertainty in the response, allowing to refer them to the second modeling stage that includes data from invasive tests. As a result, patients assessed on the first stage avoid unnecessary biopsies.

Experimental results on a publicly available breast cancer dataset exemplify the benefits of our proposal, leading to models that would have provided early non-invasive predictions to a moderate fraction of patients, without sacrificing substantial performance. Our proposal is a versatile framework that allows a wide range of uncertainty quantification functions and patient filtering patterns. This work only covered a fraction of such, and additional work should involve more extensive testing of error rates and novel non-conformity measures, e.g., using ordinal prediction sets (Lu et al., 2022).

The proposed uncertainty-aware approach has a strong potential for tackling other clinical scenario in which there might be data modalities with different associated costs. Therefore, additional research in cost-variable biomedical problems is expected in the future.

Acknowledgments

This work was partially supported by the Gobierno de Navarra through the ANDIA 2021 program (grant no. 0011-3947-2021-000023) and the ERA PerMed JTC2022 PORTRAIT project (grant no. 0011-2750-2022-000000).

References

Loai M Alnemer, Lama Rajab, and Ibrahim Aljarah. Conformal prediction technique to predict breast cancer survivability. *International Journal of Advanced Science and Technology*, 96:1–10, 2016.

Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

Elizabeth Hope Cain, Ashirbani Saha, Michael R Harowicz, Jeffrey R Marks, P Kelly Marcom, and Maciej A Mazurowski. Multivariate machine learning models for prediction of pathologic response to neoadjuvant therapy in breast cancer using mri features: a study using an independent validation set. *Breast cancer research and treatment*, 173:455–463, 2019.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26:1045–1057, 2013.

Patricia Cortazar and Charles E Geyer. Pathological complete response in neoadjuvant treatment of breast cancer. *Annals of surgical oncology*, 22:1441–1446, 2015.

Gábor Cserni, Ewa Chmielik, Bálint Cserni, and Tibor Tot. The new tnm-based staging of breast cancer. *Virchows Archiv*, 472:697–703, 2018.

Briete Goorts, Thiemo JA van Nijnatten, Linda de Munck, Martine Moossdorff, Esther M Heuts, Maaike de Boer, Marc BI Lobbes, and Marjolein L Smidt. Clinical tumor stage is the most important predictor of pathological complete response rate after neoadjuvant chemotherapy in breast cancer patients. *Breast cancer research and treatment*, 163:83–91, 2017.

Antonis Lambrou, Harris Papadopoulos, and Alex Gammerman. Evolutionary conformal prediction for breast cancer diagnosis. In *International Conference on Information Technology and Applications in Biomedicine*, pages 1–4. IEEE, 2009.

Charles Lu, Anastasios N Angelopoulos, and Stuart Pomerantz. Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pages 545–554. Springer, 2022.

Sergiusz Łukasiewicz, Marcin Czeczelewski, Alicja Forma, Jacek Baj, Robert Sitarz, and Andrzej Stanisławek. Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review. *Cancers*, 13(17):4287, 2021.

Subramani Mani, Yukun Chen, Xia Li, Lori Arlinghaus, A Bapsi Chakravarthy, Vandana Abramson, Sandeep R Bhave, Mia A Levy, Hua Xu, and Thomas E Yankeelov. Machine learning for predicting the response of breast cancer to neoadjuvant chemotherapy. *Journal of the American Medical Informatics Association*, 20(4):688–695, 2013.

Raffaella Massafra, Maria Colomba Comes, Samantha Bove, Vittorio Didonna, Gianluca Gatta, Francesco Giotta, Annarita Fanizzi, Daniele La Forgia, Agnese Latorre, Maria Irene Pastena, et al. Robustness evaluation of a deep learning model on sagittal and axial breast dce-mris to predict pathological complete response to neoadjuvant chemotherapy. *Journal of Personalized Medicine*, 12(6):953, 2022.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning*, pages 345–356. Springer, 2002.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Valeria Romeo, Giuseppe Accardo, Teresa Perillo, Luca Basso, Nunzia Garbino, Emanuele Nicolai, Simone Maurea, and Marco Salvatore. Assessment and prediction of response to neoadjuvant chemotherapy in breast cancer: A comparison of imaging modalities and future perspectives. *Cancers*, 13(14):3521, 2021.

Ashirbani Saha, Xiaozhi Yu, Dushyant Sahoo, and Maciej A Mazurowski. Effects of mri scanner parameters on breast cancer radiomics. *Expert systems with applications*, 87: 384–391, 2017.

Ashirbani Saha, Michael R Harowicz, Lars J Grimm, Connie E Kim, Sujata V Ghate, Ruth Walsh, and Maciej A Mazurowski. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 dce-mri features. *British journal of cancer*, 119(4):508–516, 2018.

Cigdem Selli and Andrew H Sims. Neoadjuvant therapy for breast cancer as a model for translational research. *Breast cancer: basic and clinical research*, 13:1178223419829072, 2019.

Laura M Spring, Geoffrey Fell, Andrea Arfe, Chandni Sharma, Rachel Greenup, Kerry L Reynolds, Barbara L Smith, Brian Alexander, Beverly Moy, Steven J Isakoff, et al. Pathologic complete response after neoadjuvant chemotherapy and impact on breast cancer recurrence and survival: A comprehensive meta-analysispcr and association with clinical outcomes in breast cancer. *Clinical cancer research*, 26(12):2838–2848, 2020.

Vladimir Vovk. Transductive conformal predictors. In *IFIP international conference on artificial intelligence applications and innovations*, pages 348–360. Springer, 2013.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2nd edition, 2022.