

PersonalizedUS: Interpretable Breast Cancer Risk Assessment with Local Coverage Uncertainty Quantification

Alek Fröhlich^{1*}, Thiago Ramos^{2*}, Gustavo Cabello³, Isabela Buzatto³, Rafael Izicki², Daniel Tiezzi³

¹UFSC, Florianópolis, Brazil

²UFSCar, São Carlos, Brazil

³USP, Ribeirão Preto, Brazil

alek.frohlich@posgrad.ufsc.br, thiagorr@ufscar.br, gustavocabelllo@usp.br, ipcarlotti@hcrp.usp.br, rizicki@ufscar.br, dtiezzi@usp.br

Abstract

Correctly assessing the malignancy of breast lesions identified during ultrasound examinations is crucial for effective clinical decision-making. However, the current “golden standard” relies on manual BI-RADS scoring by clinicians, often leading to unnecessary biopsies and a significant mental health burden on patients and their families. In this paper, we introduce PersonalizedUS, an interpretable machine learning system that leverages recent advances in conformal prediction to provide precise and personalized risk estimates with local coverage guarantees and sensitivity, specificity, and predictive values above 0.9 across various threshold levels. In particular, we identify meaningful lesion subgroups where distribution-free, model-agnostic conditional coverage holds, with approximately 90% of our prediction sets containing only the ground truth in most lesion subgroups, thus explicitly characterizing for which patients the model is most suitably applied. Moreover, we make available a curated tabular dataset of 1936 biopsied breast lesions from a recent observational multicenter study and benchmark the performance of several state-of-the-art learning algorithms. We also report a successful case study of the deployed system in the same multicenter context. Concrete clinical benefits include up to a 65% reduction in requested biopsies among BI-RADS 4a and 4b lesions, with minimal to no missed cancer cases.

Introduction

Breast cancer is the most diagnosed cancer and the leading cause of cancer-related deaths among women worldwide, with 2.3 million new cases and 666,000 deaths reported in 2022 (Bray et al. 2024). While screening mammography is widely used to increase early diagnosis—when treatment is more effective and less costly (Prager et al. 2018)—it has limitations, particularly a high false positive rate, leading to recalls, short interval follow-ups, and unnecessary biopsies (Hubbard et al. 2011). A recent study estimated that, after 10 years of annual screening in women aged 40–59 years, 61% of individuals would experience at least one false positive recall and up to 9% at least one false positive breast biopsy (Ho et al. 2022).

Breast ultrasound (US) is broadly used as a diagnostic tool complementing inconclusive mammograms, evaluating palpable findings, and guiding breast biopsies (Berg 2008).

*These authors contributed equally.

Breast US has several advantages compared to other imaging modalities, including relatively lower cost, lack of ionizing radiation, and real-time evaluation capabilities (Feig 2010).

Despite these advantages, interpreting US exams is hard. The evaluation involves inspecting image features such as lesion size, shape, and margins for signs of malignancy, and relies heavily on the operator’s experience (Sivarajah, Brown, and Chetlen 2020). Radiologists ultimately decide whether the findings are benign, need follow-up, or require a biopsy according to the Breast Imaging Reporting and Data System (BI-RADS) (Mendelson et al. 2013). Although the BI-RADS standardized the description of lesions and reports, it has notable limitations. In particular, the most recent version offers little guidance on how to subdivide BI-RADS 4 findings into subcategories (a, b, and c), with risk ranging from 5% to 95%, and overlooks clinical and Doppler-based features known to be important for malignancy prediction (Pfob et al. 2022). These gaps contribute to the method’s intra-reader variability and false positive rate (Lazarus et al. 2006; Niu et al. 2020). Notably, incorporating US in breast cancer screening leads to an additional 4–8% of patients undergoing biopsy as compared to mammography alone, yet only 7–8% of US-guided breast biopsies are found to be malignant (Yang et al. 2020; Corsetti et al. 2011).

The main goals of our work were to:

- Improve upon the current BI-RADS standard by reducing the number of false positives without compromising cancer cases;
- Provide an interpretable model that can be safely monitored by medical doctors with suitable training in AI;
- Give personalized uncertainty quantification for model predictions with strong theoretical guarantees under minimal assumptions;
- Provide an intuitive interface for doctors to interact with the model.

To address these challenges, we present a solution that integrates traditional statistical methods with modern AI techniques, with a strong emphasis on uncertainty quantification. Concretely, we developed PersonalizedUS, an interpretable and accurate logistic regression model deployed in the form of an easy-to-use web application. The key innova-

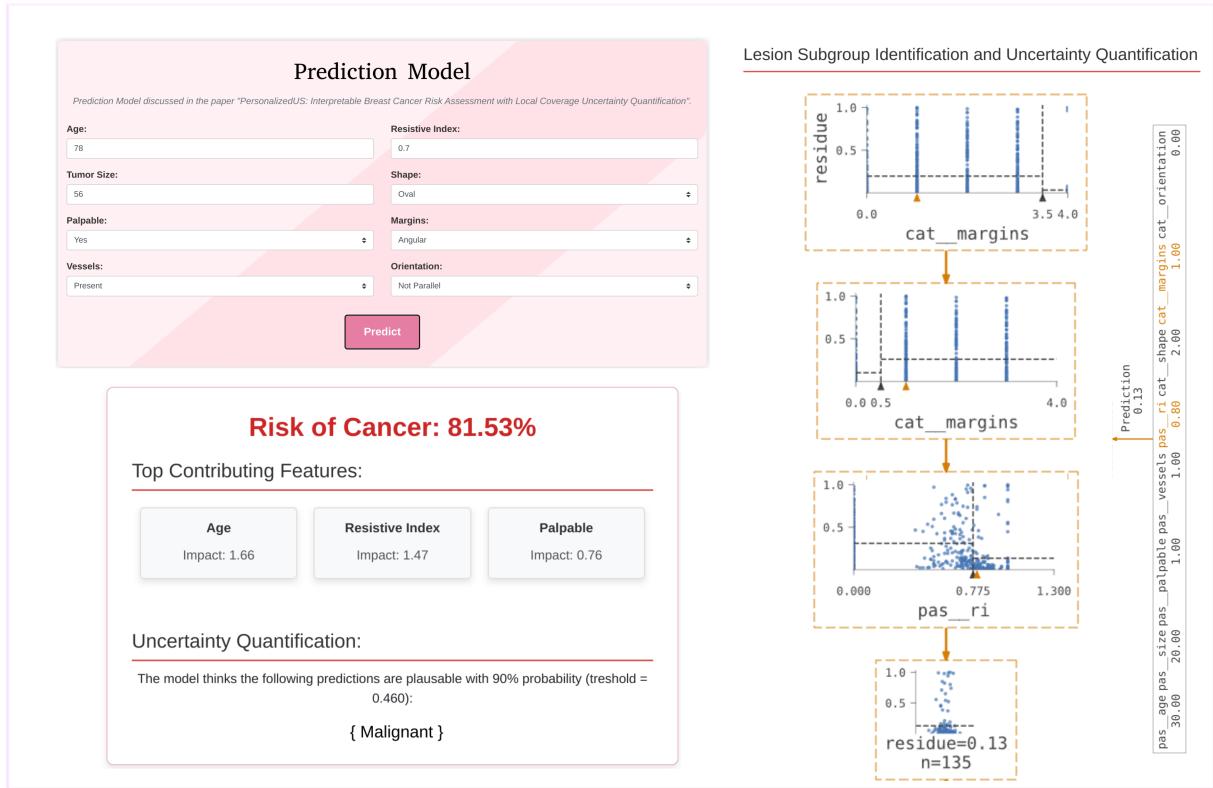


Figure 1: PersonalizedUS pipeline: Starting with a suspicious breast lesion identified by US, clinical, BI-RADS, and Doppler features are fed into a logistic regression model to estimate malignancy risk. The most influential features are highlighted. The lesion is then fed into a decision tree, whose leaves reflect difficulty levels with respect to the prediction model. Finally, a prediction set with a local conditional coverage guarantee, ensuring that lesions within more challenging leaves are associated with more uncertain predictions.

vation lies in the novel use of Locart, a recently developed conformal prediction method that leverages feature space partitions to provide personalized confidence intervals/sets that approach conditional coverage. This represents a significant advance in applying distribution-free, model-agnostic uncertainty quantification tools to the medical field, as Locart delivers localized uncertainty estimates that align more closely with the goals of precision medicine embodied by clinical prediction models.

Contributions Overview

- **PersonalizedUS:** We introduce PersonalizedUS, an interpretable machine learning system designed to accurately assess the malignancy risk of breast lesions detected by ultrasound imaging. By integrating a logistic regression model with modern conformal prediction techniques, PersonalizedUS offers precise and personalized risk estimates that improve upon the current BI-RADS standard, significantly reducing the number of false positives without missing cancer cases.
- **Novel Use of Conformal Prediction:** We adapt the Locart method, a recent approach from conformal prediction that guarantees local conditional coverage. The method was originally developed for regression problems

and is applied to binary classification for the first time here.

Related Work

There has been considerable interest in developing clinical prediction models to estimate the risk of malignancy of suspicious breast lesions identified by ultrasound. Early contributions have focused on feature extraction and classical machine learning/statistical techniques (Shen et al. 2007; Prabhakar and Poonguzhali 2017), whereas recent work has concentrated around deep neural network-based computer vision approaches (Kim et al. 2021; Qi et al. 2019; Shen et al. 2021, 2023).

Unfortunately, such attempts have often been plagued by at least one of the following issues: (i) lack of interpretability, (ii) lack of transparency/reproducibility, and (iii) lack of good software development principles. In particular, few systems go through the full software development and maintenance cycle and actually become accessible to doctors, and those that do are often based on black-box models such as deep convolutional nets (de Hond et al. 2022; Rudin 2019). Moreover, most prediction models found on the literature are either poorly developed or poorly reported (Steyerberg 2019; Collins et al. 2024).

We stress that model interpretability is a crucial, non-negotiable requirement for prediction models in healthcare (de Hond et al. 2022; Rudin 2019; Steyerberg 2019). While complex models such as deep neural nets may achieve slight improvements in predictive performance, these gains are often marginal when compared to the transparency offered by simpler models (Hastie, Tibshirani, and Friedman 2009; Doshi-Velez and Kim 2017). Logistic regression is often highlighted as a reliable and interpretable model (Steyerberg 2019), where each coefficient can be directly interpreted as the contribution of a specific variable to the risk of cancer. This is significantly more direct than existing methods for interpreting neural networks (Rudin 2019; Lipton 2018).

Interpretability is not only about understanding the model’s mechanisms but also about quantifying the confidence in its predictions. Conformal prediction (Shafer and Vovk 2008; Vovk, Gammerman, and Shafer 2005; Angelopoulos, Bates et al. 2023) is a method that provides valid measures of uncertainty for predictions in a variety of machine learning tasks, ensuring that the coverage probability of prediction intervals/sets is guaranteed under minimal assumptions.

In the field of healthcare, conformal prediction has shown significant promise (Vazquez and Facelli 2022). For instance, Csillag et al. (2023) employed conformal prediction to segment and predict the volume of amniotic fluid from fetal MRIs. Likewise, Papadopoulos, Gammerman, and Vovk (2009) utilized this method for diagnosing acute abdominal pain. Furthermore, Olsson et al. (2022) applied conformal prediction for the diagnosis and grading of prostate biopsies.

Recent advances in conformal prediction have concentrated on improving local coverage properties, tailoring predictive intervals or sets to the specific characteristics of the data (Romano, Patterson, and Candes 2019; Lei and Wasserman 2012; Izbicki, Shimizu, and Stern 2020, 2022; Cabezas et al. 2024). These enhancements make distribution-free, model-agnostic uncertainty quantification tools more closely aligned with the precision medicine objectives that clinical prediction models seek to achieve (Steyerberg 2019).

Notation

Throughout the paper we will make use of the following notation. The set $\{(X_i, Y_i)\}_{i=1}^n$ will denote an independent and identically distributed (i.i.d.) sample of pairs of breast lesion data $X_i \subset \mathcal{X} \subset \mathbb{R}^D$ and their associated malignancy $Y_i \in \{0, 1\}$ (benign or malignant, respectively). We define $[n] = \{1, \dots, n\}$. Data indices will be partitioned into $I_{\text{train}} = [n_{\text{train}}]$, $I_{\text{cal}} = [n_{\text{train}} + n_{\text{cal}}] \setminus [n_{\text{train}}]$, and $I_{\text{test}} = [n] \setminus [n_{\text{train}} + n_{\text{cal}}]$, where $n_{\text{train}} + n_{\text{cal}} + n_{\text{test}} = n$ and $n_{\text{train}}, n_{\text{cal}}, n_{\text{test}} > 0$. Any model \hat{p} will always be trained on $\{(X_i, Y_i)\}_{i \in I_{\text{train}}}$, have its output calibrated on $\{(X_i, Y_i)\}_{i \in I_{\text{cal}}}$, and evaluated on $\{(X_i, Y_i)\}_{i \in I_{\text{test}}}$.

Dataset

The Breast Lesion Dataset is comprised of $n = 1936$ breast lesions and contains both clinical, BI-RADS, and Doppler features. The lesions come from a recent multicenter observational study with a retrospective and a prospective co-

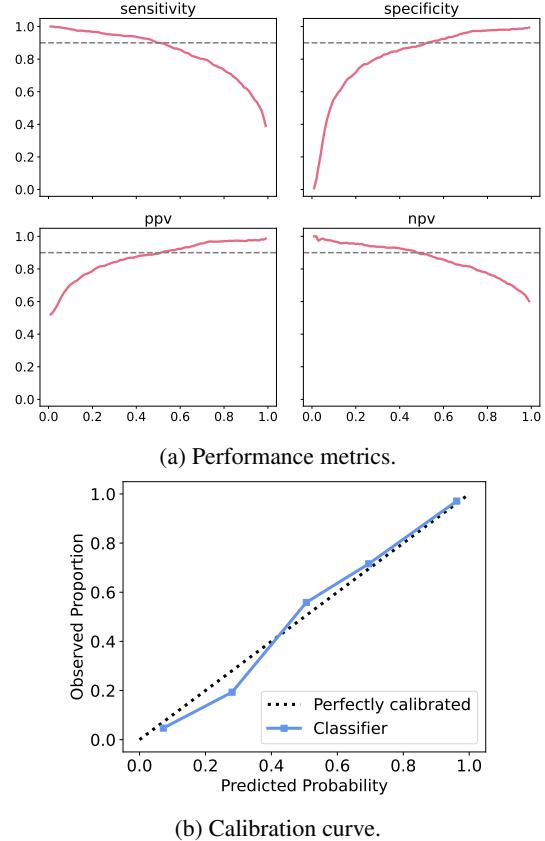


Figure 2: Results of risk estimation on the calibration set: (a) Classification metrics and (b) Calibration curve.

hort of patients (Buzatto et al. 2024). The retrospective cohort included all patients from HCRP-USP (2014-2021) and MATER (2019-2021) who underwent diagnostic ultrasound followed by core needle or excisional biopsy, whereas the prospective cohort comprised similar patients from HCRP-USP, MATER, Hospital de Amor de Barretos, and Hospital de Amor de Campo Grande, starting in 2021 and is still ongoing.

The inclusion criteria were breast lesions classified as BI-RADS at least 3 identified by ultrasound and submitted to percutaneous core needle biopsy and/or excisional biopsy. Exclusion criteria were lesions classified as BI-RADS 2, lesion size greater than 30mm, age less than 18 years old, pathologies not primary from the breast and lesions submitted exclusively to fine-needle aspiration cytology. The full details can be seen in (Buzatto et al. 2024, Methodology).

In this paper, we introduced minor modifications to the dataset. Specifically, we used a slightly updated version of the dataset and excluded entries with missing data for the variables “palpable”, “age”, and “ri” instead of performing imputations. We also retained the original BI-RADS categories of “shape” and “margins” without collapsing them. The resulting dataset was then divided into training ($n_{\text{train}} = 513$), calibration ($n_{\text{cal}} = 1059$), and test ($n_{\text{test}} = 364$) subsets. The training dataset consists solely of retrospec-

tive data, while the test dataset comprises only prospective data. The calibration dataset includes a mix of both, with its larger size being necessary to support the local conformal prediction method, which requires sufficient data per subgroup (Angelopoulos, Bates et al. 2023).

Prospective and retrospective patients have similar ages, with mean ages of 51.95 ± 13.58 and 51.44 ± 15.18 years, respectively. Both groups have similar lesion size, with means of 16.70 ± 6.99 mm in the prospective and 16.74 ± 6.85 mm in retrospective. The proportion of palpable lesions is lower in the prospective group, at 55.12%, compared to 64.17 in the retrospective. The resistance index (RI) is also slight lower in the prospective group, with a mean of 0.39 ± 0.38 , compared to 0.48 ± 0.41 in the retrospective. Irregular shapes are the most common in both groups, being more prevalent in the prospective group at 61.17% compared to 51.80%. Circumscribed margins are more frequent in the retrospective data, representing 36.51%. Orientation is mostly parallel in both groups, with a slight higher proportion in the prospective group at 73.82% compared to 67.19% in the retrospective.

A table summarizing these statistics will be provided in the supplementary material. The resulting Breast Lesion Dataset will be available as a csv file.

PersonalizedUS

We introduce PersonalizedUS, an interpretable machine learning system for assessing the malignancy of breast lesions identified by ultrasound. The system combines regularized logistic regression with modern conformal prediction tools to deliver accurate (see Table 1 and Figure 2) and interpretable predictions with personalized uncertainty quantification.

The system is deployed as an intuitive web application (see Figure 1) for specialist physicians at four Brazilian breast cancer reference centers: HCRP-USP, MATER, Hospital de Amor de Barretos, and Hospital de Amor de Campo Grande. The platform allows clinicians to input patient-specific data, receive interpretable risk assessments, and explore the detailed uncertainty measures in real-time. The interface is designed to be user-friendly, ensuring that even those with minimal technical expertise can effectively utilize the tool.

Two use cases of the system are described in “Scope and Promise for Social Impact”, along with expected social benefits in terms of facilitated access to medical resources and reduced mental health burden among women with suspected breast cancer (Dragaset and Lindstrom 2003).

A key innovation in our work is the adaptation of Locart (Cabezas et al. 2024), a conformal prediction method designed for achieving local conditional coverage, to the binary classification setting of malignancy prediction. This novel approach involves fitting a decision tree regressor to a modified calibration dataset:

$$\mathcal{D}_{\text{res}} := \{(X_i, r_i) \mid i \in I_{\text{cal}}\},$$

where r_i represent the classification residuals from the logistic model on the calibration set $\{(X_i, Y_i)\}_{i \in I_{\text{cal}}}$. The result-

ing decision tree (see Figure 4) partitions the lesion space into distinct subgroups, reflecting the model’s varying difficulty in assessing these lesions. Surprisingly, our model captures well-known patterns that expert physicians recognize in US exams, such as circumscribed lesions in young patients, spiculated lesions with high vascularization, and other less-defined subgroups that are more challenging for doctors to manage.

Locart calibrates the risk assessment model locally within each partition using standard conformal prediction techniques (Angelopoulos, Bates et al. 2023). This approach generates personalized confidence sets for each prediction, giving clinicians both a risk estimate and a clear measure of uncertainty. As demonstrated by Theorems 1 and 2, and shown empirically in Table 2, the method achieves conditional coverage, unlike standard conformal prediction methods that only provide marginalized coverage.

The source code for PersonalizedUS, along with its documentation, will be published in a public repository after the anonymous review process. The web application will also be accessible following the review.

Methodology

Risk Assessment Model

In our study, we trained a logistic regression model to assess the malignancy of breast lesions identified during ultrasound examinations. Logistic regression was chosen for its simplicity and interpretability, making it highly suitable for healthcare applications (Steyerberg 2019).

The model was developed on the training set following a grid search procedure based on 5-fold cross validation for optimizing the regularization parameter $C \in \{0.01, 0.1, \dots, 100\}$. The model with the lowest log-loss was chosen as the best model. The model was implemented in scikit-learn, with the categorical features “margins” and “shape” being one-hot encoded and the numerical features “age”, “size”, and “ri” being standardized.

Baselines

To provide a robust evaluation of our logistic regression model, we compared its performance against several other widely-used machine learning algorithms, including decision trees, Naive Bayes, k-nearest neighbors (KNN), XG-Boost, neural networks, AdaBoost, random forest, and support vector machines (SVM). Each model was trained and evaluated under the same experimental conditions to ensure a fair comparison.

The results are summarized in Table 1. Our logistic regression model achieved a competitive performance, with a strong balance between sensitivity (0.900) and specificity (0.888). Although some more complex models showed slightly higher metrics, the differences were marginal. These results suggest that there are no significant advantages in using more complex models at the cost of the interpretability offered by logistic regression.

Model	Sensitivity	Specificity	PPV	NPV
Logistic	0.900	0.888	0.897	0.892
DT	0.770	0.947	0.940	0.793
NB	0.687	0.975	0.967	0.743
KNN	0.903	0.894	0.902	0.896
XGB	0.865	0.871	0.878	0.857
NN	0.885	0.904	0.908	0.880
AB	0.818	0.908	0.905	0.822
RF	0.900	0.878	0.888	0.891
SVM	0.898	0.896	0.903	0.891

Table 1: Comparison of learning algorithms over the Breast Lesion Dataset.

Personalized Uncertainty Quantification

In this section, we present a novel method to construct interpretable and adaptive prediction sets for the predictions of our risk model. Concretely, we adapt the Locart method to the binary classification setting of malignancy prediction (Cabezas et al. 2024).

The method is designed to partition the lesion space in a way that reflects the varying difficulty of malignancy prediction with respect to our machine learning model. Standard conformal prediction methods are then applied to these local partitions, thereby providing local coverage guarantees for individual predictions. The details of this approach are outlined next.

1. **Calculating residuals.** Let \hat{p} be the malignancy prediction model. For each $i \in I_{\text{cal}}$, we began by calculating the residuals over the calibration set:

$$r_i = \begin{cases} 1 - \hat{p}(Y = 1 | X_i), & \text{if } y_i = 1, \\ 1 - \hat{p}(Y = 0 | X_i), & \text{if } y_i = 0. \end{cases}$$

2. **Creating residual dataset.** A new dataset \mathcal{D}_{res} was then created by combining the calibration data \mathcal{D}_{cal} with the residuals r_i , that is:

$$\mathcal{D}_{\text{res}} := \{(X_i, r_i) \mid i \in I_{\text{cal}}\}.$$

Then, we split this dataset in two subsets: $\mathcal{D}_{\text{res},0}$ and $\mathcal{D}_{\text{res},1}$, such that, $\mathcal{D}_{\text{res}} = \mathcal{D}_{\text{res},0} \cup \mathcal{D}_{\text{res},1}$.

3. **Learning patient subgroups.** We trained a decision tree regressor \hat{T} over $\mathcal{D}_{\text{res},0}$. This decision tree induced a partition of the feature space with size K :

$$\mathcal{X} = \bigcup_{i=1}^K \mathcal{X}_i.$$

We expect this partition to effectively capture and reflect the varying complexity of cancer predictions within each lesion subgroup, ensuring that the risk model's local behavior is accurately represented.

4. **Calculating quantiles for each subgroup.** Given a miscoverage level $\alpha \in (0, 1)$, for each partition learned from $\mathcal{D}_{\text{res},0}$, we calculated an adjusted coverage quantile $\tilde{\alpha}$

over $\mathcal{D}_{\text{res},1}$. Specifically, for each element \mathcal{X}_j of the partition, the quantile q_j was determined as:

$$q_{j,1-\alpha} = \text{Quantile}_{1-\tilde{\alpha}}(r_i : (X_i, r_i) \in \mathcal{D}_{\text{res},1}, X_i \in \mathcal{X}_j),$$

where

$$\tilde{\alpha} = \frac{\lceil (k_j + 1) \cdot \alpha \rceil}{|k_j|},$$

is the adjusted significance level of $\alpha \in (0, 1)$ (Angelopoulos, Bates et al. 2023) and

$$k_j := |\{(X_i, r_i) \in \mathcal{D}_{\text{res},1} \mid X_i \in \mathcal{X}_j\}|$$

is the number of patients in $\mathcal{D}_{\text{res},1}$ that are also in \mathcal{X}_j .

5. **Personalized quantification of uncertainty.** Given a test lesion X_i , $i \in I_{\text{test}}$, we check which element \mathcal{X}_j of the partition it belongs, then we define its prediction set as

$$C_{\text{locart}}(x) = \{\ell \in \{0, 1\} : \hat{p}(y = \ell | X = x) \geq 1 - q_{j,1-\alpha}\}.$$

The theoretical foundations for this method are established in (Cabezas et al. 2024) and explained next.

Theorem 1 (Theorem 2 in Cabezas et al. (2024)). *Let $\{\mathcal{X}_1, \dots, \mathcal{X}_K\}$ be a finite partition of the lesion space. Given an exchangeable sequence $(X_i, y_i)_{i=1}^{n+1}$ and a miscoverage level $\alpha \in (0, 1)$, the following holds:*

$$\mathbb{P}[Y_{n+1} \in C_{\text{locart}}(X_{n+1}) \mid X_{n+1} \in \mathcal{X}_j] \geq 1 - \tilde{\alpha},$$

for all $j = 1, \dots, K$ where $\tilde{\alpha} = \frac{\lceil (k_j + 1) \cdot \alpha \rceil}{|k_j|}$ and $k_j := |\{i \in [n] \mid X_i \in \mathcal{X}_j\}|$.

The theorem guarantees that, given a finite partition of the feature space, the probability that the prediction for a new sample falls within the local prediction band—conditioned on the sample belonging to one of these partitions—is at least $1 - \alpha$.

The core idea behind the in this approach model is that, by leveraging the partition learned from the decision tree, we can achieve local conditional coverage. In fact, Cabezas et al. (2024) show that this local method for uncertainty quantification achieves asymptotic coverage.

Theorem 2. *Given a lesion $x \in \mathcal{X}$, let $\mathcal{X}(x)$ be the element of the learned partition given by Locart which x belongs to. Under certain regularity conditions, Locart has conditional validity, that is*

$$\lim_{n \rightarrow \infty} \mathbb{P}[Y_{n+1} \in C_{\text{locart}}(X_{n+1}) \mid X_{n+1} \in \mathcal{X}(x)] = \mathbb{P}[Y_{n+1} \in C(X_{n+1}) \mid X_{n+1} = x],$$

where $C(\cdot)$ is the theoretical optimal confidence set.

In summary, integrating Locart into PersonalizedUS marks a significant step forward in uncertainty quantification for breast cancer risk assessment. This approach ensures that each prediction comes with a locally adaptive confidence interval, enhancing both the reliability and interpretability of the model, ultimately making it a more effective tool for clinical decision-making.

Learning Lesion Subgroups

Following the steps outlined in the previous section, we fitted a decision tree regressor to the residual dataset. The tree model was developed on the basis of a grid search procedure via 5-fold cross-validation for optimizing the structural parameters “max_depth” $\in [3, 4, 5, 6]$ and “min_samples_leaf” $\in [70, 80, 90, 100]$. The model with the lowest mean squared error was chosen as the best model. The model was implemented in `scikit-learn`, with the categorical features “shape” and “margins” being ordinal-encoded.

Evaluation

Risk Estimation

The selected logistic regression model was then evaluated on the calibration set to determine its performance. We compared the model’s risk estimates to the actual outcomes and calculated key performance metrics, including sensitivity (recall of class 1), specificity (recall of class 0), positive predictive value (precision of class 1), and negative predictive value (precision of class 0), the resulting curves can be seen in Figure 2 along with the model’s calibration curve. We also evaluated the Area Under the ROC curve (0.958), the Area Under the Precision-Recall Curve (0.960), the log-loss (0.268).

Lesion Subgroup Evaluation

Two medical doctors with specialized training in AI (IB and DT) evaluated the resulting tree structure depicted in Figure 4. According to their expert assessment, the model successfully captures patterns that physicians commonly recognize in ultrasound exams, such as circumscribed lesions in young patients, spiculated lesions with high vascularization, and other less-defined subgroups that are more challenging to manage.

As illustrated in Figure 3, these patterns are also evident in the distribution of BI-RADS categories and malignancy across subgroups. The leftmost groups predominantly feature benign BI-RADS 3 and 4a lesions, while the rightmost groups contain most BI-RADS 4c, 5, and malignant lesions, which are generally easier to predict. The middle groups, however, are more mixed, consisting mainly of BI-RADS 4 and 5 lesions with varying levels of malignancy. These lesions proved especially challenging for our logistic regression model.

Uncertainty Quantification Evaluation

Prediction sets generated by Locart were evaluated according to three criteria: (i) average set size, (ii) comparison between theoretical and empirical coverage, and (iii) proportion of prediction sets containing only the ground truth. A successful application of conformal prediction to binary classification is expected to often produce singleton sets, with empirical coverage close to $1 - \alpha$. A successful application of Locart further requires that these metrics hold conditionally to $X_j \in \mathcal{X}_J$ for $j \in I_{\text{test}}$. As such, in Table 2 we show the result of applying the three evaluation criteria

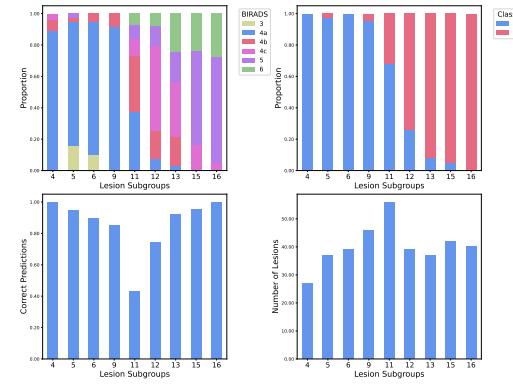


Figure 3: Lesion subgroups analysis of BI-RADS categories, malignancy, predictive accuracy, and subgroup size.

with $\alpha = 0.1$, grouping by leaf. Also, we show a stacked plot of set sizes in Figure 5.

As we can see from the first column of Table 2 and Figure 5, the average set size was close to 1 for most leaves, except for leaves 11 and 12. This pattern is also evident in the last column of Table 2, where lesions within leaves 11 and 12 were considerably more challenging for our risk model to assess. This behavior is expected, as Locart groups lesions based on their relative difficulty for the malignancy prediction model. Therefore, leaves 11 and 12 contain the more challenging lesions, enabling tighter predictive sets in the remaining leaves. Notably, these two “hard” leaves represent only 20% of the test set, suggesting that our model reliably predicts malignancy for 80% of the dataset, which, like the entire dataset, consists solely of suspicious lesions that were biopsied. Finally, the observed fluctuation of up to 10% in empirical coverage falls within the expected range, given the sample size of approximately 100 points per leaf (Angelopoulos, Bates et al. 2023).

Leaf	Avg. Set Size	Emp. Cov. (%)	Truth Only (%)
4	0.96	96.43	96.43
5	0.97	94.59	94.59
6	0.97	97.44	97.44
9	1.00	95.65	95.65
11	1.55	89.29	33.93
12	1.41	94.87	53.85
13	0.97	89.19	89.19
15	0.93	90.48	90.48
16	1.00	100.00	100.00

Table 2: Evaluation of our local conformal prediction tool.

Scope and Promise for Social Impact

Two significant social issues arise from false positive US evaluations: (i) an increased burden on the healthcare system, particularly on pathologists, which delays timely access to treatment for those in need, and (ii) the substantial

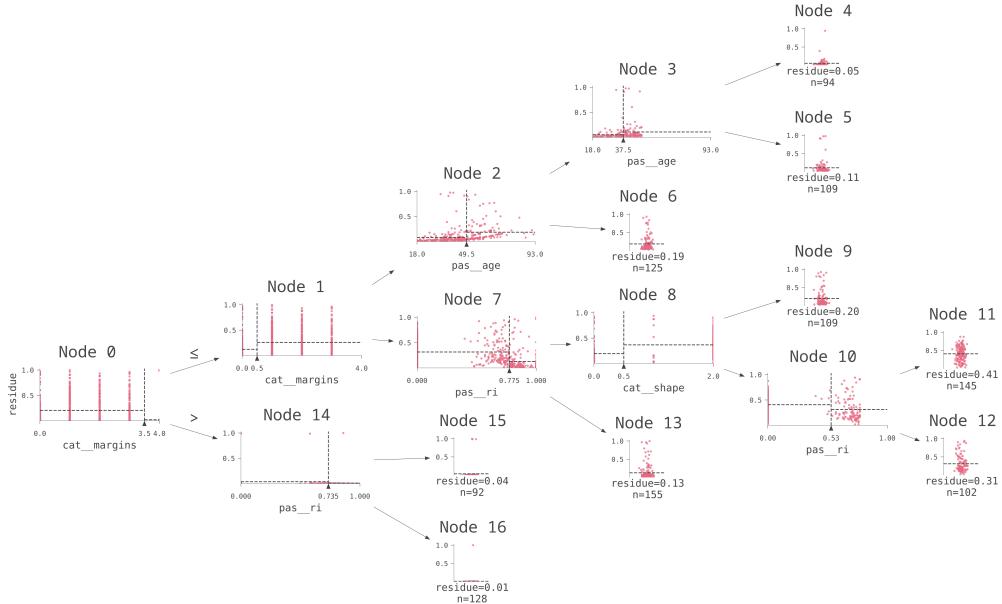


Figure 4: Decision tree regressor trained for predicting the classification residuals of our logistic regression model over the calibration set.

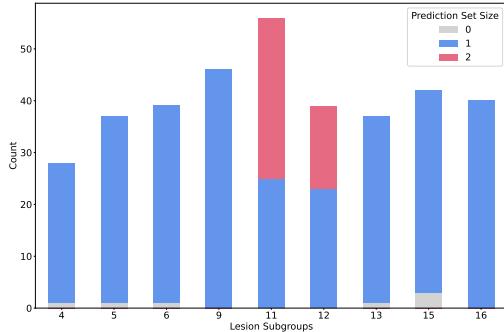


Figure 5: Stacked plot of (local) average set sizes.

mental health impact on patients and their families, especially in underserved communities (Dragset and Lindstrom 2003). This section discusses two potential applications of our prediction model aimed at optimizing the biopsy referral process to ensure it prioritizes those who really need it, and alleviate the stress and anxiety experienced by patients and their families.

The first, most straightforward application of our model is to assist radiologists in differentiating between benign and malignant lesions. In this scenario, the physician would initially evaluate the US exam and then, specially for lesions classified as BI-RADS 4a and 4b, use our machine learning system for a second opinion, if desired. In cases where the model is highly confident the lesion is benign, the radiologist would be prompted to carefully reconsider whether a biopsy referral is necessary.

Another, less intrusive application of our prediction model is to rank patients awaiting biopsy based on their as-

sessed risks as determined by our model. Even though this approach wouldn't reduce the overall load on the healthcare system, it would help ensure that high-risk patients receive priority access to medical resources.

To evaluate the effectiveness of our system in differentiating between benign and malignant breast masses, we further analysed its test set performance by considering only BI-RADS 4a and 4b lesions. Using an approach similar to that of Buzatto et al. (2024), we selected a threshold that maximized the negative predictive value while ensuring the positive predictive value did not fall below that of BI-RADS 4b (Yoon et al. 2011). If the model's recommendations were followed, only 64 out of 196 biopsies would be requested, representing a reduction of more than 65%. Moreover, with the optimized threshold ($t = 0.223$), the model wouldn't miss any cancer cases among BI-RADS 4a and 4b lesions.

Conclusion and Follow-up Work

This paper introduces PersonalizedUS, an interpretable AI system for predicting malignancy of breast lesions identified by US. The system combines logistic regression with new conformal prediction methods to provide accurate predictions with individualized uncertainty quantification.

The system is deployed as a web application for specialists at four Brazilian breast cancer reference centers. The site allows doctors to enter patient data, receive interpretable risk estimates, and explore uncertainty quantification. We explored two use cases of the system, along with expected social benefits in terms of facilitated access to medical resources and reduced mental health burden among women with suspected breast cancer.

A novelty in our work is the use of Locart, a conformal prediction method designed for achieving local conditional

coverage. This method partitions the lesion space, reflecting the model's varying difficulty in assessing these lesions. Our tree captures well-known patterns that specialists recognize in US exams, such as circumscribed lesions in young patients, spiculated lesions with high vascularization, and other less-defined subgroups that are harder for doctors to manage.

Facilitation of Follow-up Work Locart's model-agnostic nature allows our logistic regression model to be easily replaced by more complex models, such as neural networks, without affecting the personalized uncertainty quantification. Also, we will make both our data and code publicly available after the review process.

References

- Angelopoulos, A. N.; Bates, S.; et al. 2023. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4): 494–591.
- Berg, W. A. 2008. Combined Screening With Ultrasound and Mammography vs Mammography Alone in Women at Elevated Risk of Breast Cancer. *JAMA*, 299(18): 2151–2163.
- Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R. L.; Soerjomataram, I.; and Jemal, A. 2024. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*.
- Buzatto, I. P. C.; Recife, S. A.; Miguel, L.; Bonini, R. M.; Onari, N.; Faim, A. L. P. A.; Silvestre, L.; Carlotti, D. P.; Fröhlich, A.; and Tiezzi, D. G. 2024. Machine learning can reliably predict malignancy of breast lesions based on clinical and ultrasonographic features. *Breast Cancer Research and Treatment*.
- Cabezas, L.; Otto, M. P.; Izbicki, R.; and Stern, R. B. 2024. Regression Trees for Fast and Adaptive Prediction Intervals. *arXiv preprint arXiv:2402.07357*.
- Collins, G. S.; Moons, K. G. M.; Dhiman, P.; Riley, R. D.; Beam, A. L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J. B.; van Smeden, M.; Boulesteix, A.-L.; Camaradou, J. C.; Celi, L. A.; Denaxas, S.; Denniston, A. K.; Glocker, B.; Golub, R. M.; Harvey, H.; Heinze, G.; Hoffman, M. M.; Kengne, A. P.; Lam, E.; Lee, N.; Loder, E. W.; Maier-Hein, L.; Mateen, B. A.; McCradden, M. D.; Oakden-Rayner, L.; Ordish, J.; Parnell, R.; Rose, S.; Singh, K.; Wynants, L.; and Logullo, P. 2024. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, e078378.
- Corsetti, V.; Houssami, N.; Ghirardi, M.; Ferrari, A.; Speziani, M.; Bellarosa, S.; Remida, G.; Gasparotti, C.; Gallegioni, E.; and Ciatto, S. 2011. Evidence of the effect of adjunct ultrasound screening in women with mammography-negative dense breasts: Interval breast cancers at 1-year follow-up. *European Journal of Cancer*, 47(7): 1021–1026.
- Csillag, D.; Paes, L. M.; Ramos, T.; Romano, J. V.; Schuller, R.; de Beauclair Seixas, R.; Oliveira, R. I.; and Orenstein, P. 2023. AmnioML: Amniotic Fluid Segmentation and Volume Prediction with Uncertainty Quantification. In Williams, B.; Chen, Y.; and Neville, J., eds., *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7–14, 2023*, 15494–15502. AAAI Press.
- de Hond, A. A. H.; Leeuwenberg, A. M.; Hooft, L.; Kant, I. M. J.; Nijman, S. W. J.; van Os, H. J. A.; Aardoom, J. J.; Debray, T. P. A.; Schuit, E.; van Smeden, M.; Reitsma, J. B.; Steyerberg, E. W.; Chavannes, N. H.; and Moons, K. G. M. 2022. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *npj Digital Medicine*, 5(1).
- Doshi-Velez, F.; and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- Dragaset, S.; and Lindstrom, T. C. 2003. The mental health of women with suspected breast cancer: the relationship between social support, anxiety, coping and defence in maintaining mental health. *Journal of Psychiatric and Mental Health Nursing*, 10(4): 401–409.
- Feig, S. 2010. Cost-Effectiveness of Mammography, MRI, and Ultrasonography for Breast Cancer Screening. *Radioologic Clinics of North America*, 48(5): 879–891.
- Hastie, T.; Tibshirani, R.; and Friedman, J. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
- Ho, T.-Q. H.; Bissell, M. C. S.; Kerlikowske, K.; Hubbard, R. A.; Sprague, B. L.; Lee, C. I.; Tice, J. A.; Tosteson, A. N. A.; and Miglioretti, D. L. 2022. Cumulative Probability of False-Positive Results After 10 Years of Screening With Digital Breast Tomosynthesis vs Digital Mammography. *JAMA Network Open*, 5(3): e222440.
- Hubbard, R. A.; Kerlikowske, K.; Flowers, C. I.; Yankaskas, B. C.; Zhu, W.; and Miglioretti, D. L. 2011. Cumulative Probability of False-Positive Recall or Biopsy Recommendation After 10 Years of Screening Mammography: A Cohort Study. *Annals of Internal Medicine*, 155(8): 481.
- Izbicki, R.; Shimizu, G.; and Stern, R. 2020. Flexible distribution-free conditional predictive bands using density estimators. In *International Conference on Artificial Intelligence and Statistics*, 3068–3077. PMLR.
- Izbicki, R.; Shimizu, G.; and Stern, R. B. 2022. Cd-split and hpd-split: Efficient conformal regions in high dimensions. *Journal of Machine Learning Research*, 23(87): 1–32.
- Kim, S.-Y.; Choi, Y.; Kim, E.-K.; Han, B.-K.; Yoon, J. H.; Choi, J. S.; and Chang, J. M. 2021. Deep learning-based computer-aided diagnosis in screening breast ultrasound to reduce false-positive diagnoses. *Scientific Reports*, 11(395).
- Lazarus, E.; Mainiero, M. B.; Schepps, B.; Koelliker, S. L.; and Livingston, L. S. 2006. BI-RADS Lexicon for US and Mammography: Interobserver Variability and Positive Predictive Value. *Radiology*, 239(2): 385–391.
- Lei, J.; and Wasserman, L. 2012. Distribution Free Prediction Bands. *arXiv:1203.5422*.
- Lipton, Z. C. 2018. The mythos of model interpretability. *Communications of the ACM*, 61(10): 36–43.

- Mendelson, E.; Böhm-Vélez, M.; Berg, W.; et al. 2013. *ACR BI-RADS Ultrasound*. American College of Radiology.
- Niu, S.; Huang, J.; Li, J.; Liu, X.; Wang, D.; Zhang, R.; Wang, Y.; Shen, H.; Qi, M.; Xiao, Y.; Guan, M.; Liu, H.; Li, D.; Liu, F.; Wang, X.; Xiong, Y.; Gao, S.; Wang, X.; and Zhu, J. 2020. Application of ultrasound artificial intelligence in the differential diagnosis between benign and malignant breast lesions of BI-RADS 4A. *BMC Cancer*, 20(1).
- Olsson, H.; Kartasalo, K.; Mulliqi, N.; Capuccini, M.; Ruusuvuori, P.; Samaratunga, H.; Delahunt, B.; Lindskog, C.; Janssen, E.; Blilie, A.; ISUP Prostate Imagebase Expert Panel; Egevad, L.; Spjuth, O.; and Eklund, M. 2022. Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. *Nature Communications*, 13(1).
- Papadopoulos, H.; Gammerman, A.; and Vovk, V. 2009. Reliable diagnosis of acute abdominal pain with conformal prediction. *International journal of engineering intelligent systems for electrical engineering and communications*, 17: 127–137.
- Pfob, A.; Sidey-Gibbons, C.; Barr, R. G.; Duda, V.; Alwafai, Z.; Balleymguier, C.; Clevert, D.-A.; Fastner, S.; Gomez, C.; Goncalo, M.; Gruber, I.; Hahn, M.; Hennigs, A.; Kapetas, P.; Lu, S.-C.; Nees, J.; Ohlinger, R.; Riedel, F.; Rutten, M.; Schaeffgen, B.; Schuessler, M.; Stieber, A.; Togawa, R.; Tozaki, M.; Wojcinski, S.; Xu, C.; Rauch, G.; Heil, J.; and Golatta, M. 2022. The importance of multi-modal imaging and clinical information for humans and AI-based algorithms to classify breast masses (INSPiRED 003): an international, multicenter analysis. *European Radiology*, 32(6): 4101–4115.
- Prabhakar, T.; and Poonguzhal, S. 2017. Automatic detection and classification of benign and malignant lesions in breast ultrasound images using texture morphological and fractal features. *2017 10th Biomedical Engineering International Conference (BMEiCON)*, 1–5.
- Prager, G. W.; Braga, S.; Bystricky, B.; Qvortrup, C.; Criscitiello, C.; Esin, E.; Sonke, G. S.; Martinez, G.; Frenel, J.-S.; Karamouzis, M.; Strijbos, M.; Yazici, O.; Bossi, P.; Banerjee, S.; Troiani, T.; Eniu, A.; Ciardiello, F.; Tabernero, J.; Zielinski, C. C.; Casali, P. G.; Cardoso, F.; Douillard, J.-Y.; Jezdic, S.; McGregor, K.; Bricalli, G.; Vyas, M.; and Ilbawi, A. 2018. Global cancer control: responding to the growing burden, rising costs and inequalities in access. *ESMO Open*, 3(2): e000285.
- Qi, X.; Zhang, L.; Chen, Y.; Pi, Y.; Chen, Y.; Lv, Q.; and Yi, Z. 2019. Automated diagnosis of breast ultrasonography images using deep neural networks. *Medical Image Analysis*, 52: 185–198.
- Romano, Y.; Patterson, E.; and Candes, E. 2019. Conformalized Quantile Regression. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215.
- Shafer, G.; and Vovk, V. 2008. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9(3).
- Shen, W.-C.; Chang, R.-F.; Moon, W. K.; Chou, Y.-H.; and Huang, C.-S. 2007. Breast Ultrasound Computer-Aided Diagnosis Using BI-RADS Features. *Academic Radiology*, 14(8): 928–939.
- Shen, Y.; Park, J.; Yeung, F.; Goldberg, E.; Heacock, L.; Shamout, F.; and Geras, K. J. 2023. Leveraging Transformers to Improve Breast Cancer Classification and Risk Assessment with Multi-modal and Longitudinal Data. arXiv:2311.03217.
- Shen, Y.; Shamout, F. E.; Oliver, J. R.; Witowski, J.; Kannan, K.; Park, J.; Wu, N.; Huddleston, C.; Wolfson, S.; Milliet, A.; Ehrenpreis, R.; Awal, D.; Tyma, C.; Samreen, N.; Gao, Y.; Chhor, C.; Gandhi, S.; Lee, C.; Kumari-Subaiya, S.; Leonard, C.; Mohammed, R.; Moczulski, C.; Altabet, J.; Babb, J.; Lewin, A.; Reig, B.; Moy, L.; Heacock, L.; and Geras, K. J. 2021. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. *Nature Communications*, 12(1).
- Sivarajah, R. T.; Brown, K.; and Chetlen, A. 2020. “I can see clearly now.” fundamentals of breast ultrasound optimization. *Clinical Imaging*, 64: 124–135.
- Steyerberg, E. W. 2019. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer International Publishing. ISBN 9783030163990.
- Vazquez, J.; and Facelli, J. C. 2022. Conformal Prediction in Clinical Medical Sciences. *Journal of healthcare informatics research*, 6(3): 241—252.
- Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic learning in a random world*, volume 29. Springer.
- Yang, L.; Wang, S.; Zhang, L.; Sheng, C.; Song, F.; Wang, P.; and Huang, Y. 2020. Performance of ultrasonography screening for breast cancer: a systematic review and meta-analysis. *BMC Cancer*, 20(1).
- Yoon, J. H.; Kim, M. J.; Moon, H. J.; Kwak, J. Y.; and Kim, E.-K. 2011. Subcategorization of Ultrasonographic BI-RADS Category 4: Positive Predictive Value and Clinical Factors Affecting It. *Ultrasound in Medicine & Biology*, 37(5): 693–699.