End-to-End Learning of Fused Image and Non-Image Features for Improved Breast Cancer Classification from MRI

Gregory Holste[1] Savannah C. Partridge[2,3] Habib Rahbar[2,3] Debosmita Biswas[3]
Christoph I. Lee[2,3] Adam M. Alessio[1*]

[1]Michigan State University [2]University of Washington [3]Seattle Cancer Care Alliance

[*]aalessio@msu.edu

Abstract

Breast cancer diagnosis is inherently multimodal. To assess a patient's cancer status, physicians integrate imaging findings with a variety of clinical risk factor data. Despite this, deep learning approaches for automatic breast cancer classification often only utilize image data or non-image clinical data, but not both simultaneously. In this work, we implemented and compared strategies for the fusion of imaging and tabular non-image data in an end-to-end trainable manner, evaluating fusion at different stages in the model (fusing intermediate features vs. output probabilities) and with different operations (concatenation vs. addition vs. multiplication). This retrospective study utilized dynamic contrast-enhanced MRI (DCE-MRI) data from 10,185 breast MRI examinations of 5,248 women. DCE-MRIs were reduced to 2D maximum intensity projections, split into single-breast images, then linked to a set of 18 non-image features including clinical indication and mammographic breast density. We first trained unimodal baseline models on images alone and non-image data alone. We then developed three multimodal fusion models that learn jointly from image and non-image data, evaluating performance by area under the receiver operating characteristic curve (AUC) and specificity at 95% sensitivity. The image-only baseline achieved an AUC of 0.849 (95% CI: 0.834, 0.864) and specificity at 95% sensitivity of 30.1% (95% CI: 23.1%, 37.0%), while the best-performing fusion model achieved an AUC of 0.898 (95% CI: 0.885, 0.909) and specificity of 49.1% (95% CI: 38.8%, 55.3%). Furthermore, all three fusion methods significantly outperformed both unimodal baselines with respect to AUC and specificity at 95% sensitivity. This work demonstrates in our dataset for breast cancer classification that incorporating non-image data with images can significantly improve predictive performance and that fusion of intermediate learned features is superior to fusion of final probabilities.

1. Introduction

In breast imaging, radiologists use heterogeneous information from multiple sources to decide whether and to what extent a patient exhibits risk for breast cancer [18, 16]. The information available to radiologists may include imaging findings, demographic and clinical data (age, gender, and clinical indication), and information on cancer risk factors (comorbidities, family history, and breast density [16]). Despite this, deep learning approaches to breast cancer diagnosis typically rely only on patient imaging or only on patient risk factors. Based on the assumption that imaging features and clinical features offer independent diagnostic value, a model that fuses and learns jointly from these different types of information may increase predictive performance over a unimodal (single-source) approach.

Deep convolutional neural networks (CNNs) have been successfully applied to a wide variety of tasks in medical diagnostics, particularly for detecting breast cancer from screening mammograms [17, 24, 23, 3] and diagnostic MRI studies [12, 29, 30]. However, these efforts rarely take advantage of clinical risk factor data that is readily available to the interpreting radiologist or metadata associated with patient imaging. Some studies have created "multimodal" inputs to CNNs by combining images of different modalities [9, 27, 28] in an "early fusion" fashion. For the late fusion approach, other have combined representations from different imaging modalities [15, 31, 26] and others have learned from both image and tabular patient data, but only in a post-hoc manner that combines information from independently trained models (i.e., an ensemble or multi-stage model) [6, 1, 30, 22]. Recent work proposed an end-to-end fusion method that combines histology images with genomic profiles for glioma survival prediction, observing significant improvement over an image-only CNN baseline by concordance index (from 0.75 to 0.826) [7]. Likewise, the authors in [1] fused information from mammography with clinical data in a two-stage fashion, seeing similar improvements over a CNN baseline by AUC (from 0.88 to 0.91).

These prior works, however, either (a) neglected an exploration of different strategies for merging heterogeneous information or (b) developed fusion methods that rely on combining independently trained models. Instead, we seek a single architecture that can be optimized end-to-end to learn jointly from image and non-image data.

In this work, we explore strategies for fusing information from patient imaging with associated non-image data *in an end-to-end trainable manner*, hypothesizing that non-image features can add predictive value to a CNN that otherwise could only accommodate images. As a use case, we train and evaluate these methods on dynamic contrast-enhanced MRI (DCE-MRI) images and associated clinical data for the task of identifying breast cancer. We propose three intuitive fusion methods and conduct a feature importance analysis to illustrate which non-image features are most critical to automatic breast cancer prediction; we follow this up with additional experiments to uncover the best way to fuse information from two modalities and the best way to optimize such a network.

2. Materials and Methods

2.1. Data Collection and Description

In this retrospective, institutional review board-approved study (Fred Hutchinson Cancer Research Center protocol #7339), we curated fully anonymized data from 5,248 women who underwent 10,185 breast cancer examinations at the University of Washington from July 2005-November 2015. Each patient received a DCE-MRI exam, 76.5% of patients also received a mammogram, and 26.8% underwent a breast tissue biopsy. If a patient had a pathology-confirmed breast cancer at the time of examination or received a cancer diagnosis within 12 months after MRI, that breast was labeled "Malignant;" all other breasts were labeled "Benign." Additional features such as patient age, clinical indication for MRI (e.g., high risk screening, diagnostic evaluation, assess extent of disease, or other), and background parenchymal enhancement (BPE) from MRI were collected as well. Each DCE-MRI data set was reduced to a 2D maximum intensity projection (MIP) image of the peak contrast enhanced minus background image using vendor software. MRI performed post-biopsy could potentially have susceptibility artifacts or other signal changes from biopsy clips, presenting the biasing effect that clip artifacts could be strongly associated with malignancy. The use of MIP images reduces this bias, as clip artifacts are generally not visible in MIPs that only capture areas of contrast enhancement.

Since breast cancer diagnosis is a breast-specific task, we consider each breast to be an individual case for the purposes of model training and evaluation. We first ensured that all breasts in our study cohort had a Breast Imaging-Reporting and Data System (BI-RADS) assessment of 1-6 (information not used for training) and were given a breast cancer status within 12 months of MRI examination. We then removed all breasts from studies with corrupted image files or observed artifacts in the DCE-MRI; artifacts occurred mostly due to failure in the MIP generation process, as evidenced by high intensity regions visually not consistent with vascular patterns. From this set of 17,046 breasts from over 5,000 women, we then processed the MIPs as described below, linking them to a set of non-image features to create a multimodal data set of imaging and tabular clinical information.

2.2. Data Preparation

After DCE-MRI data sets were reduced to 2D MIP images by vendor software, MATLAB (Mathworks, Natick, MA) was used to preprocess MIPs for model training and evaluation. Each image was cropped to split the MIP into two single-breast images, and then the top 0.5% of pixel intensities of each image were clipped, assuming these extreme values represented noise due to artifacts. Lastly, each image was resized to 224×224 pixels and its intensity values were linearly normalized to the interval $[0, 1]$. Alongside these image processing steps, basic information from the images – such as the total width and height (in *mm*) and MRI system software version – were appended to the list of non-image features. These steps produced a final data set of 17,046 breast images, each with an associated vector of 18 tabular features (see Supplemental Materials for full description). We denote these tabular features, containing clinical and MRI acquisition information, as "non-image" features because they are not radiomic features derived from the image itself.

Next, continuous-valued non-image features were standardized, and categorical and ordinal features were dichotomized into binary variables via dummy coding; this produced a total of 33 non-image inputs. To learn from a sufficient proportion of malignant breasts, we obfuscated non-image features that were directly linked to the final breast cancer status. While BI-RADS assessments were not used as inputs, breasts with a BI-RADS 6 assessment are known to have cancer prior to examination, meaning the patient's "Assess Extent of Disease" indication would be highly correlated with a malignant outcome; for BI-RADS 6 cases, we replaced the indication from "Assess Extent of Disease" uniformly at random with one of the other three indications: "Screening," "Diagnostic," or "Other." This effectively changed the indication feature such that the remaining "Assess Extent of Disease" cases were those in which we observed a previous positive biopsy for cancer in the contralateral breast. Missing values of age, BPE, and mammographic breast density (if mammography not performed) were then imputed with the median age and most

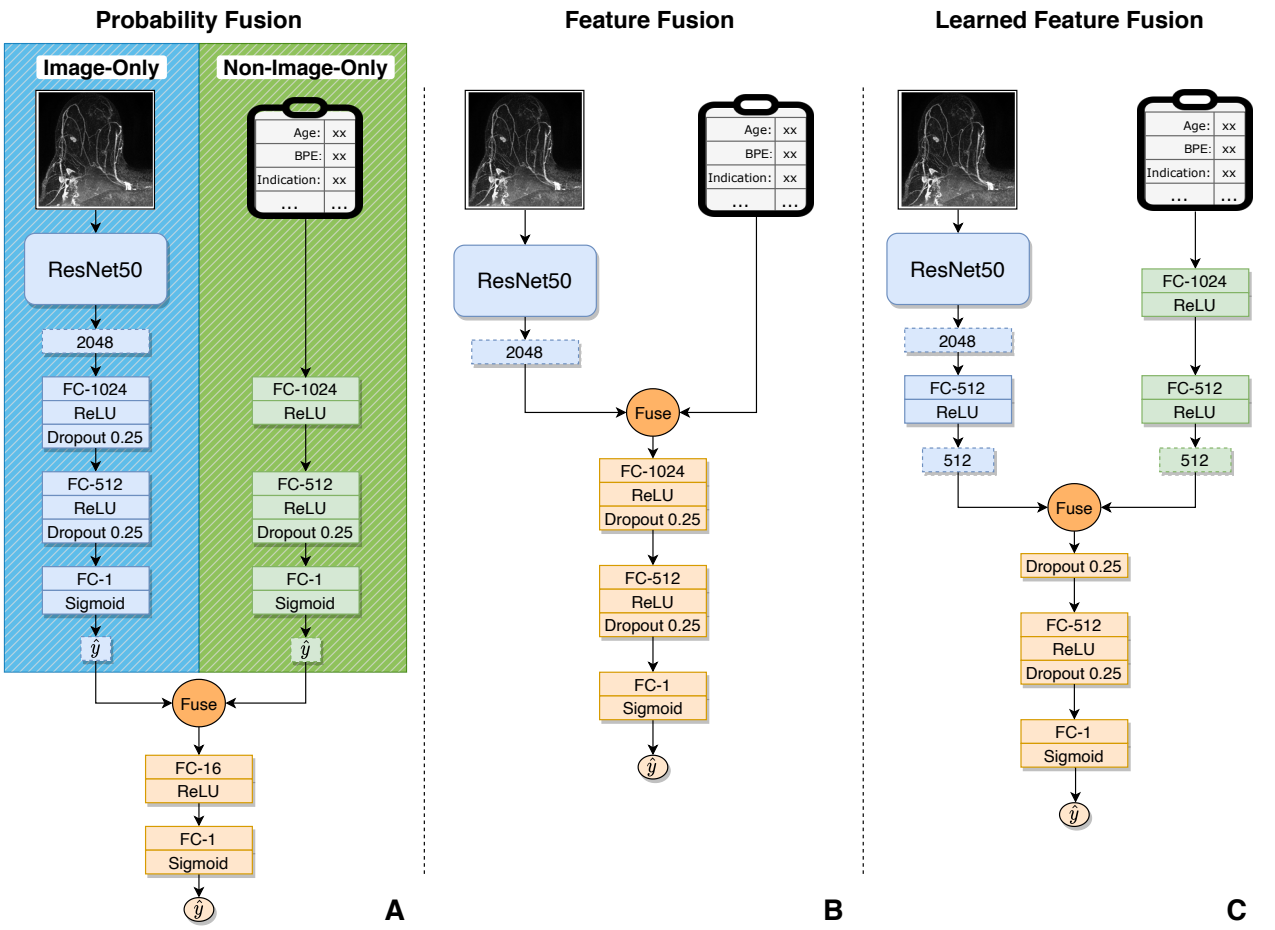

Figure 1: Diagram of fusion architectures that learn jointly from breast imaging and tabular non-image features. *Probability Fusion* (A) fuses information at the output level; *Feature Fusion* (B) fuses learned image features with non-image features; *Learned Feature Fusion* (C) fuses learned image features with learned non-image features. The baseline *Image-Only* and *Non-Image-Only* models are seen, respectively, in the blue and green regions of panel A; that is, the unimodal baseline architectures are components of the *Probability Fusion* architecture. Dashed boxes represent feature vectors, with the number inside representing the size of that vector. "FC-n" represents a fully-connected layer with n hidden units. The symbol $\hat{y}$ represents a predicted probability of malignancy within the next 12 months.

frequently observed breast density and BPE in the training set. Finally, we randomly assigned 10,466 cases for training (61.4%), 1,671 cases for validation to mitigate overfitting (9.8%), and 4,909 cases for testing (28.8%), ensuring that each patient appeared in only one of the three sets.

2.3. Deep Learning Model Architectures

To evaluate if fusion of image and non-image features can improve breast cancer prediction, we first established a baseline image-only and non-image feature-only approach. Our *Image-Only* model consisted of ResNet50 [10], adapted to accommodate a single-channel input, with two fully-connected layers followed by a single output neuron replacing the original classification head. Our *Non-Image-Only* model was a simple feedforward neural network with two fully-connected layers followed by a single output neuron (Figure 1A).

We explore three primary approaches to fuse image-dervied features with tabular non-image features, varying at what stage in the multimodal architecture features are fused. The first, *Probability Fusion*, considers the output probabilities of an image-only and non-image-only model as inputs to a fully-connected layer that produces a final prediction; observe that *Probability Fusion* contains the baseline *Image-Only* and *Non-Image-Only* models within itself (Figure 1A). *Feature Fusion* learns a vector of 2,048 features from a breast image, and concatenates the 33 non-image inputs onto that extracted feature vector before learning jointly from image and non-image features to produce a final prediction (Figure 1B). Lastly, *Learned Feature Fu-*

sion simultaneously learns features from the breast image and non-image features, then concatenates learned feature vectors from each modality before learning from this combined vector to produce a final prediction (Figure 1C). The latter two methods allow for the model to directly learn interactions between image and non-image features, while the former only fuses information at the prediction level.

2.4. Model Nomenclature

In one survey of multimodal machine learning techniques [4], the authors provide a taxonomy for fusion approaches, which we find to be limited in expressive power. For example, what we call *Feature Fusion* and *Learned Feature Fusion* would be indistinguishable under their framework – they would both be categorized as "model-agnostic, early fusion." A more recent review of techniques specifically for fusing medical imaging with tabular clinical data [13] provides a more comprehensive naming scheme; under their system, *Probability Fusion* would be called "Late Fusion," *Feature Fusion* would be called "Joint Fusion – Type II," and *Learned Feature Fusion* would be called "Joint Fusion – Type I." While expressive enough to capture the approaches examined here, these names do not communicate what exactly is being fused or *how* features from different modalities are being fused (concatenation, averaging, *etc.*). To propose a flexible naming system for "late fusion" methods, we first observe that there are three types of features one could combine during training: probabilities (P), learned features (L), and semantic features (S). Using brackets to represent concatenation, *Probability Fusion* could be called *[P,P]-Fusion* (concatenating probabilities from the two modalities), *Feature Fusion* could be called *[L,S]-Fusion* (concatenating learned image features with semantic non-image features), and *Learned Feature Fusion* could be called *[L,L]-Fusion* (concatenating learned image features with learned non-image features).

2.5. Experiments Varying Fusion Operation

While the main fusion experiments use concatenation to fuse information from different modalities, the architectures presented in Figure 1 are general enough to accomodate other fusion operations. To understand how the fusion operation impacts predictive performance, we also trained variants of the *Learned Feature Fusion* model that, instead of concatenating features from each modality, elementwise add (called L+*L-Fusion*) and multiply (called $L\times$*L-Fusion*) feature vectors. A full description of auxiliary experiments involving different ways to train this multimodal architecture (optimizing each subnetwork – image encoder, non-image encoder, fusion head – separately vs. optimizing the entire network with a single loss expression) can be found in the Supplemental Materials.

Table 1. Characteristics across training, validation, and test sets.

	Training Set	Validation Set	Test Set
Cases	10,466	1,671	4,909
Age (*yr*)*	51.5 ± 11.1	51.9 ± 10.7	51.5 ± 11.5
Laterality			
Left	5,229 (50.0%)	842 (50.4%)	2,476 (50.4%)
Right	5,237 (50.0%)	829 (49.6%)	2,433 (49.6%)
MRI Indication			
Other	1,075 (10.3%)	175 (10.5%)	557 (11.3%)
Screening	6,906 (66.0%)	1,150 (68.8%)	3,186 (64.9%)
Diagnostic	1,135 (10.8%)	152 (9.1%)	566 (11.5%)
Assess extent of disease	1,350 (12.9%)	194 (11.6%)	600 (12.2%)
Breast Density			
Entirely fatty	207 (2.0%)	28 (1.7%)	116 (2.4%)
Scattered density	2,157 (20.6%)	374 (22.4%)	880 (18.0%)
Heterogeneously dense	6,953 (66.4%)	1,071 (64.1%)	3,285 (67.0%)
Extremely dense	1,149 (11.0%)	198 (11.8%)	628 (12.8%)
BPE			
Minimal	6,132 (58.6%)	1,002 (60.0%)	2,819 (57.4%)
Mild	2,563 (24.5%)	382 (22.9%)	1,241 (25.3%)
Moderate	1,169 (11.2%)	208 (12.4%)	561 (11.4%)
Marked	602 (5.8%)	79 (4.7%)	288 (5.9%)
Positive Biopsy†	2,090 (20.0%)	325 (19.4%)	1,006 (20.5%)

A "case" is a single-breast MIP image and associated vector of non-image features. Unless indicated otherwise, values represent numbers of cases and values in parentheses represent percentages out of the total number of cases in a given set.

* Values represent mean ± standard deviation.

† Denotes a biopsy-proven malignancy within 12 months post-MRI.

2.6. Experimental Setup and Analysis

We trained the five architectures described in Section 2.3 and the two variants described in Section 2.5 to predict breast cancer status, comparing the performance of fusion models to the baseline *Image-Only* and *Non-Image-Only* models by (a) AUC and (b) specificity at 95% sensitivity, when evaluated on the test set. All models were trained with randomly initialized weights under identical training regimes: the same optimizer, learning rate, data augmentations, early stopping schedule, *etc.* (see Supplemental Materials for details). Models were created and trained in PyTorch version 1.4.0 [19].

We used the pROC package [21] in R version 4.0.0 [20] to compute nonparametric confidence intervals and conduct significance tests used to assess model performance. All confidence intervals were obtained with 5,000 stratified bootstrap samples of the test set via the percentile method. A simple nonparametric test for difference in means (see Supplemental Materials for details) was used to determine the significance of differences in model performance; this method was chosen over the popular DeLong test so that it could be applied to metrics other than AUC. A P-value less than 0.05 was selected to demonstrate a statistically significant effect.

Lastly, for the five trained models described in Section

Table 2. Breast cancer prediction results of multimodal fusion models and their unimodal baselines.

Model	Image Inputs?	Non-Image Inputs?	Best Run		Five-Run Ensemble	
			AUC	Specificity at 95% Sensitivity (%)	AUC	Specificity at 95% Sensitivity (%)
Image-Only	✓		0.849 [0.834, 0.864]	30.1 [23.1, 37.0]	0.860 [0.845, 0.873]	33.2 [27.4, 39.7]
Non-Image-Only		✓	0.807 [0.791, 0.823]	27.7 [22.3, 33.9]	0.806 [0.790, 0.821]	29.5 [20.5, 34.4]
Probability Fusion	✓	✓	0.888 [0.875, 0.899]	48.2 [42.2, 53.5]	0.888 [0.876, 0.899]	51.3 [45.2, 56.2]
Feature Fusion	✓	✓	0.894 [0.882, 0.905]	46.5 [40.5, 51.1]	0.901 [0.890, 0.913]	47.3 [42.6, 55.0]
Learned Feature Fusion	✓	✓	0.898 [0.885, 0.909]	49.1 [38.8, 55.3]	0.903 [0.891, 0.914]	50.3 [44.2, 59.0]

Values represent the specified performance metric, and values in brackets represent 95% bootstrapped confidence intervals obtained on the test set (N=4,909). Each model was trained five separate times; "Best Run" refers to the single model realization with maximum validation AUC, and "Five-Run Ensemble" refers to an ensemble of the five realizations of each model.

2.3, we conducted a feature importance analysis to assess which non-image features were most influential to each model. We used a permutation-based measure of performance (as first described in [5]), where we randomly shuffled the values of a single feature in the test set, generated new predictions for the permuted test data, and recalculated AUC. The "importance" of that feature was then the percent reduction in test AUC upon permutation – the intuition being that permuting an important feature will result in an appreciable change in model performance (see Supplementary Materials for details). Code will be available at https://github.com/gholste/breast_mri_fusion.

3. Results

3.1. Study Cohort

The training, validation, and test sets consisted of cases – single-breast (unilateral) images with associated non-image features – with roughly equal composition with respect to age, clinical indication, mammographic breast density, and biopsy-confirmed cancer status (Table 1). The training set contained 10,466 cases from 3,015 women (aged 51.5 $\pm$ 11.1 years [mean $\pm$ standard deviation]), roughly 20% of cases having biopsy-confirmed breast cancer within a year of examination. About 77.4% of training images contained dense breasts, as observed on mammography (BI-RADS C or D), and 83.1% exhibited low BPE (minimal or mild). Further, after obscuring patients' "Assess Extent of Disease" status for the ipsilateral breast as described earlier, about 13% of cases came from patients with known cancer in the contralateral breast.

3.2. Model Results

All five deep learning architectures were trained end-to-end for a maximum of 100 epochs, restoring model weights from the epoch with the highest AUC on the validation set. Additionally, each model was trained five separate times with different random number-generating seeds so that we could consider ensembles of each model across five unique random weight initializations ("runs"). The run with highest maximum validation AUC is used for all analysis presented in the text, but ensemble results can be found in Tables 2 and 4.

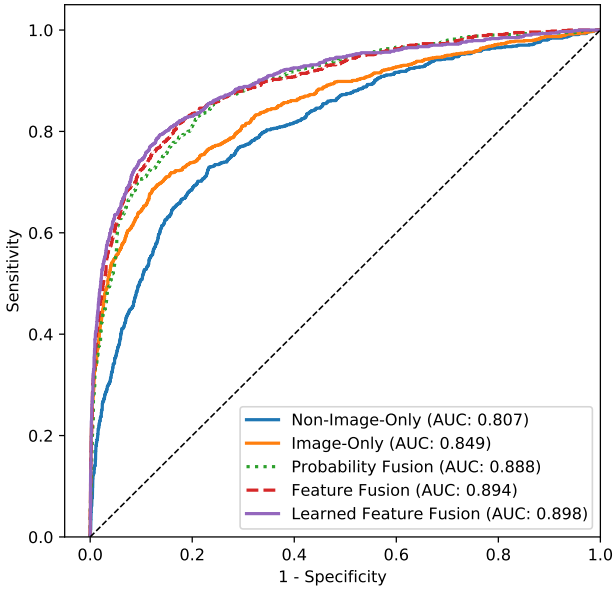

Figure 2: Comparison of model performance on breast cancer prediction for all five model architectures considered. *Non-Image-Only* was only trained on tabular clinical features and *Image-Only* was only trained on breast images, while the remaining fusion models were trained on both images and clinical features. Fusion models significantly outperformed Image-Only and Shallow-Only with respect to AUC ($P \ll 0.001$ for each test). AUC = area under the receiver operating characteristic curve.

Overall, models trained on both patient imaging and associated non-image features outperform their image-only and non-image feature-only counterparts by both AUC and specificity at 95% sensitivity (Table 2, Figure 2). The baseline *Image-Only* model achieved an AUC of 0.849 (95% CI: 0.834, 0.64), while the *Non-Image-Only* model achieved an

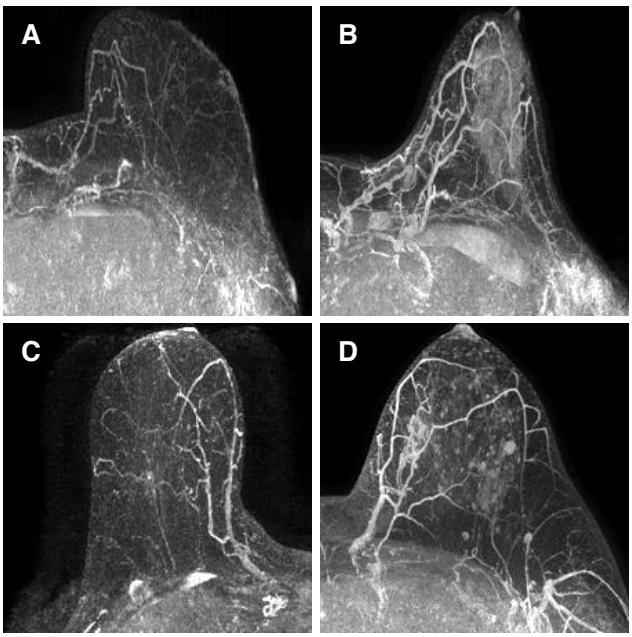

Figure 3: Example true negative (A), false positive (B), false negative (C), and true positive (D) cases according to the Image-Only model, where the "positive" class represents a pathology-confirmed cancer diagnosis in that particular breast within 12 months of MRI. Each image is a fully preprocessed MIP of the DCE-MRI study, and predictions were made at an operating threshold of 0.5. While the unimodal Image-Only model misclassified the images in panels B and C, every fusion model reversed these errors, correctly classifying image B as benign and image C as malignant with the aid of shallow clinical features. MIP = maximum intensity projection, DCE-MRI = dynamic contrast-enhanced MRI.

AUC of 0.807 (95% CI: 0.791, 0.823). As for fusion models trained on both images and tabular features, *Probability Fusion* achieved an AUC of 0.888 (95% CI: 0.875, 0.899), *Feature Fusion* achieved an AUC of 0.894 (95% CI: 0.882, 0.905), and *Learned Feature Fusion* achieved an AUC of 0.898 (95% CI: 0.885, 0.909). We also find that, at a highly sensitive operating point, fusion approaches decrease the false positive rate when compared to unimodal baselines; namely, specificity at 95% sensitivity improves from 30.1% (95% CI: 23.1%, 37.0%) in the *Image-Only* model to as high as 49.1% (95% CI: 38.8%, 55.3%) in the *Learned Feature Fusion* model. Each fusion approach significantly outperformed the *Image-Only* baseline with respect to AUC ($P \ll 0.001$ for each test) and specificity at 95% sensitivity ($P < 0.01$ for each test).

To offer additional context to the clinically relevant goal of specificity at a high sensitivity, the *Learned Feature Fusion* model correctly identifies 743 more benign cases (out of the total 3,903 present in the test set) than the *Image-Only* baseline, representing a 63% increase in specificity. Furthermore, at a given threshold, both false positive and false negative predictions by the Image-Only model are often corrected by fusion methods (Figure 3). Though the *Image-Only* model misclassified the breast in Figure 3B as malignant ($\hat{y}$ = 0.834 probability of malignancy), all fusion methods were able to remedy this error: *Probability Fusion* producing $\hat{y}$ = 0.434, *Feature Fusion* $\hat{y}$ = 0.340, and *Learned Feature Fusion* $\hat{y}$ = 0.252. Similarly, the *Image-Only* model misclassified the breast in Figure 3C as benign ($\hat{y}$ = 0.438), but with the context that this patient was 77 years old with a diagnostic indication and dense breasts according to previous mammography, all fusion methods output an increased predicted probability of cancer: *Probability Fusion* producing $\hat{y}$ = 0.691, *Feature Fusion* $\hat{y}$ = 0.742, and *Learned Feature Fusion* $\hat{y}$ = 0.956.

Comparing fusion approaches to one another, *Learned Feature Fusion* is the best-performing model with respect to AUC and specificity at 95% sensitivity. *Learned Feature Fusion* significantly outperforms *Probability Fusion* by AUC ($P = 0.003$) but does not significantly outperform *Feature Fusion* ($P = 0.247$). In summary, the two models that merge intermediate features (before the output and after the input level) outperform the model that merges output

Table 3. Feature importance summary of top non-image features.

Non-Image Feature	**Permutation Importance (Rank)**			
	Non-Image-Only	*Probability Fusion*	*Feature Fusion*	*Learned Feature Fusion*
MRI Indication	21.26 (1)	8.29 (1)	5.72 (1)	7.17 (1)
MRI Software Version	2.71 (2)	1.14 (2)	0.35 (4)	0.96 (3)
Age	2.55 (3)	0.64 (5)	0.93 (2)	1.74 (2)
MIP Height	1.81 (5)	0.33 (6)	0.55 (3)	0.80 (4)
Precession Frequency	1.01 (8)	1.11 (3)	0.31 (5)	0.31 (7)
MIP Max Intensity	2.50 (4)	0.21 (8)	0.00 (18)	0.31 (8)
BPE	1.33 (7)	0.14 (10)	0.23 (6)	0.11 (13)
Reconstruction Diameter	0.13 (15)	0.31 (7)	0.03 (13)	0.44 (5)
Pixel Dimensions	1.60 (6)	0.83 (4)	-0.00 (17)	-0.05 (17)
Repetition Time	0.69 (10)	0.11 (11)	0.18 (7)	0.15 (10)
Breast Density	0.45 (13)	0.16 (9)	0.16 (8)	0.29 (9)

Values represent median percent reduction in AUC ("permutation importance") upon permuting the values of the specified feature in the test set 30 times. Values in parentheses represent the rank of the specified feature with respect to absolute permutation importance relative to all 18 non-image features for the specified model. Features are sorted by increasing rank product, the geometric mean of each feature's importance ranks across the four models trained on shallow features. A full description of all 18 non-image features can be found in the Supplemental Materials. MIP = maximum intensity projection, AUC = area under the receiver operating characteristic curve.

Table 4. Breast cancer prediction results of *Learned Feature Fusion* model with different fusion operations.

Model	Best Run		Five-Run Ensemble	
	AUC	**Specificity at 95% Sensitivity (%)**	**AUC**	**Specificity at 95% Sensitivity (%)**
[L,L]-Fusion	0.898 [0.885, 0.909]	49.1 [38.8, 55.3]	0.903 [0.891, 0.914]	50.3 [44.2, 59.0]
L+L-Fusion	0.895 [0.883, 0.906]	49.0 [41.3, 55.4]	0.902 [0.890, 0.913]	49.1 [40.7, 57.3]
L×L-Fusion	0.893 [0.881, 0.905]	50.8 [45.7, 55.0]	0.896 [0.884, 0.907]	50.3 [42.8, 56.9]

Values represent the specified performance metric, and values in brackets represent 95% bootstrapped confidence intervals obtained on the test set (N=4,909). Each model was trained five separate times; "Best Run" refers to the single model realization with maximum validation AUC, and "Five-Run Ensemble" refers to an ensemble of the five realizations of each model.

probabilities before learning a final decision. There were insufficient data to demonstrate a significant difference with respect to specificity at 95% sensitivity ($P > 0.5$ for each test), suggesting that all fusion models were comparably specific at a very sensitive operating point.

3.3. Feature Importance Analysis

To obtain a permutation-based measure of feature importance, we permuted the values of each feature 30 times to find the resulting decrease in test set AUC upon permutation. Applying this method to the four models that used non-image features yields the feature importance ranking seen in Table 3. Overall, we found that clinical indication, MRI system software version, and age were the most important variables across all four models that were trained on non-image data. The *Non-Image-Only* model, in particular, relies heavily on clinical indication and displays generally greater absolute importance values than the three fusion models for virtually every feature. Additionally, we found that BPE from the DCE-MRI was ranked higher in importance than mammographic density by rank product, which is consistent with recent evidence that BPE is more strongly associated with breast cancer risk than mammographic breast density [2, 8].

3.4. Fusion Operation

While all models presented above fused features by concatenation, one can integrate information from multiple modalities with a variety of other operations. Results from variants of the *Learned Feature Fusion* model that elementwise added (*L+L-Fusion*) and elementwise multiplied (*L×L-Fusion*) learned features can be found in Table 4. These models performed comparably with the concatenation-based version of *Learned Feature Fusion*, and have the advantage of slightly fewer learnable parameters. While the single-run performance metrics of *L+L-Fusion* and *L×L-Fusion* did not reach those of the original *[L,L]-Fusion*, there was not enough evidence to conclude a significant difference in performance by AUC or specificity at 95% sensitivity ($P > 0.05$ for each test).

4. Discussion

Utilizing a large breast MRI database, we demonstrated that multimodal fusion models that learn jointly from breast images and non-image features significantly outperformed image-only and non-image feature-only models for automated breast cancer prediction. Furthermore, fusion models that allowed for interactions between learned image features and non-image features outperformed the approach of combining output probabilities from an image-only and non-image-only model. Therefore, our results suggest that incorporating readily available metadata associated with patient imaging can significantly improve predictive performance in deep-learned approaches for automated diagnostics.

A feature importance analysis revealed that clinical indication, patient age, and MRI version were the most salient features for breast cancer prediction in both non-image-only models and fusion models that learned from images as well. Furthermore, other than knowledge of clinical indication, the relative importance ranking of non-image features varied across the different models, suggesting that some non-image features have information that can be partially learned directly from the images; this follows intuition that some non-image features may be highly correlated with image-derived features, offering less predictive value in fusion models. This, in addition to the fact that non-image features are simply not "competing" with image features in a unimodal model, may explain why the *Non-Image-Only* model exhibits greater absolute importance values than fusion models for nearly every feature. We recognize that our finding that MRI software version, independent of images, is related to malignancy status is not a clinically relevant finding and will not translate to other datasets and clinical environments. We present the results here in an effort to show that there can be confounding effects in the data, likely due to biased data collection.

While other studies have combined image data with clin-

ical information – even utilizing richer data such as lab results and genomic profiles – they often fuse these data sources by combining separately trained models [33, 14] or by training a model in multiple stages [6, 1] (not end-to-end). Some studies have proposed end-to-end solutions that fuse image and non-image features [7, 11, 32, 25], seeing similar improvements in predictive performance, but these approaches can be unintuitive or neglect the multitude of ways in which one could combine information from different modalities. Our main contribution is the examination of three straightforward methods of fusing features from different modalities, as well as follow-up experiments to understand the best operations to join information from different modalities. Among end-to-end fusion approaches for the task of breast cancer classification, we found that fusing output probabilities was inferior to fusing intermediate learned features from each modality. Furthermore, we find that there is no apparent performance difference between combining learned image features with learned non-image features and combining learned image features with the raw semantic non-image features. Lastly, supplementary experiments revealed that optimizing such a multimodal architecture in ways other than computing a single cross-entropy loss term produces competitive results, especially upon ensembling.

While models were trained with a large multimodal data set, this study was limited due to its retrospective design based on data from a single institution; future work would be needed to evaluate if our trained models generalize to data from other sites. Likewise, non-image features were missing from about 2.3% of all cases, requiring imputation that may confound results. As for model architectures and training, in order to compare models as fairly as possible, we kept all hyperparameters fixed and only explored a small fraction of possible variations in model design. As with any deep learning study, more optimal hyperparameter tuning and architecture design choices may produce slightly different results than those presented here. Lastly, this study used a single 2D MIP image to summarize a multi-series breast MRI acquisition. Models that could incorporate volumetric MRI data or additional sequences from the breast MRI exam would likely further increase predictive performance.

5. Conclusion

In conclusion, we examined three general deep learning approaches for combining information from breast MRI studies with non-image clinical features, finding that each fusion approach significantly outperformed an image-only and non-image-only model for breast cancer prediction. These results suggest that researchers intending to train a deep learning system on patient imaging should consider what additional information is already available to them, including standard demographic and risk factor data collected at the time of breast imaging. Even with basic non-image information, one may be able to significantly improve predictive performance with one of the fusion approaches presented here.

Acknowledgments

The authors would like to thank Muneeza Azmat and Carina Pereira for their initial efforts in the data processing for this study. This work was supported by NIH/NCI CCSG P30 CA015704, NIH/NCI R37 CA240403, and NIH/NICHD R21 HD097609.

References

[1] Ayelet Akselrod-Ballin, Michal Chorev, Yoel Shoshan, Adam Spiro, Alon Hazan, Roie Melamed, Ella Barkan, Esma Herzel, Shaked Naor, Ehud Karavani, Gideon Koren, Yaara Goldschmidt, Varda Shalev, Michal Rosen-Zvi, and Michal Guindy. Predicting Breast Cancer by Applying Deep Learning to Linked Health Records and Mammograms. *Radiology*, 292(2):331–342, June 2019. Publisher: Radiological Society of North America. 1, 8

[2] Vignesh A. Arasu, Diana L. Miglioretti, Brian L. Sprague, Nila H. Alsheik, Diana S.M. Buist, Louise M. Henderson, Sally D. Herschorn, Janie M. Lee, Tracy Onega, Garth H. Rauscher, Karen J. Wernli, Constance D. Lehman, and Karla Kerlikowske. Population-Based Assessment of the Association Between Magnetic Resonance Imaging Background Parenchymal Enhancement and Future Primary Breast Cancer Risk. *J Clin Oncol*, 37(12):954–963, Apr. 2019. 7

[3] John Arevalo, Fabio A González, Raúl Ramos-Pollán, Jose L Oliveira, and Miguel Angel Guevara Lopez. Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer methods and programs in biomedicine*, 127:248–257, 2016. 1

[4] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443, Feb. 2019. 4

[5] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. Publisher: Springer. 5

[6] Anika Cheerla and Olivier Gevaert. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics*, 35(14):i446–i454, 2019. 1, 8

[7] R. J. Chen, M. Y. Lu, J. Wang, D. F. K. Williamson, S. J. Rodig, N. I. Lindeman, and F. Mahmood. Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. *IEEE Transactions on Medical Imaging*, pages 1–1, 2020. Conference Name: IEEE Transactions on Medical Imaging. 1, 8

[8] Brian N. Dontchos, Habib Rahbar, Savannah C. Partridge, Larissa A. Korde, Diana L. Lam, John R. Scheel, Sue Peacock, and Constance D. Lehman. Are Qualitative Assessments of Background Parenchymal Enhancement, Amount of Fibroglandular Tissue on MR Images, and Mammo-

graphic Density Associated with Breast Cancer Risk? *Radiology*, 276(2):371–380, Aug. 2015. 7

[9] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li. Deep Learning-Based Image Segmentation on Multimodal Medical Imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 3(2):162–169, Mar. 2019. Conference Name: IEEE Transactions on Radiation and Plasma Medical Sciences. 1

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. ISSN: 1063-6919. 3

[11] Seok-Jae Heo, Yangwook Kim, Sehyun Yun, Sung-Shil Lim, Jihyun Kim, Chung-Mo Nam, Eun-Cheol Park, Inkyung Jung, and Jin-Ha Yoon. Deep Learning Algorithms with Demographic Information Help to Detect Tuberculosis in Chest Radiographs in Annual Workers' Health Examination Data. *Int J Environ Res Public Health*, 16(2), 2019. 8

[12] Qiyuan Hu, Heather M. Whitney, and Maryellen L. Giger. A deep learning methodology for improved breast cancer diagnosis using multiparametric MRI. *Sci Rep*, 10(1):10536, 2020. 1

[13] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P. Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digital Medicine*, 3(1):1–9, Oct. 2020. Number: 1 Publisher: Nature Publishing Group. 4

[14] Shih-Cheng Huang, Anuj Pareek, Roham Zamanian, Imon Banerjee, and Matthew P. Lungren. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific Reports*, 10(1):22147, Dec. 2020. Number: 1 Publisher: Nature Publishing Group. 8

[15] Zhong Liu, Shaobin Zhong, Qiang Liu, Chenxi Xie, Yunzhu Dai, Chuan Peng, Xin Chen, and Ruhai Zou. Thyroid nodule recognition using a joint convolutional neural network with information fusion of ultrasound images and radiofrequency data. *European Radiology*, 31(7):5001–5011, 2021. 1

[16] Valerie A. McCormack and Isabel dos Santos Silva. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. *Cancer Epidemiol Biomarkers Prev*, 15(6):1159–1169, June 2006. 1

[17] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S. Corrado, Ara Darzi, et al. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94, Jan. 2020. Number: 7788 Publisher: Nature Publishing Group. 1

[18] Heidi D. Nelson, Bernadette Zakher, Amy Cantor, Rongwei Fu, Jessica Griffin, Ellen S. O'Meara, Diana S.M. Buist, Karla Kerlikowske, Nicolien T. van Ravesteyn, Amy Trentham-Dietz, Jeanne Mandelblatt, and Diana Miglioretti. Risk Factors for Breast Cancer for Women Age 40 to 49: A Systematic Review and Meta-analysis. *Ann Intern Med*, 156(9):635–648, May 2012. 1

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019. 4

[20] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. 4

[21] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. pROC: An Open-source Package for R and S+ to Analyze and Compare ROC Curves. *BMC Bioinformatics*, 12(1):77, Mar. 2011. 4

[22] Ritabrata Sanyal, Devroop Kar, and Ram Sarkar. Carcinoma type classification from high-resolution breast microscopy images using a hybrid ensemble of deep convolutional features and gradient boosting trees classifiers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021. 1

[23] Thomas Schaffter, Diana S. M. Buist, Christoph I. Lee, Yaroslav Nikulin, Dezso Ribli, Yuanfang Guan, William Lotter, Zequn Jie, Hao Du, Sijia Wang, et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. *JAMA Netw Open*, 3(3):e200265, Mar. 2020. 1

[24] Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Scientific Reports*, 9(1):12495, Aug. 2019. Number: 1 Publisher: Nature Publishing Group. 1

[25] Simeon E. Spasov, Luca Passamonti, Andrea Duggento, Pietro Lio, and Nicola Toschi. A Multi-modal Convolutional Neural Network Framework for the Prediction of Alzheimer's Disease. *Annu Int Conf IEEE Eng Med Biol Soc*, 2018:1271–1274, July 2018. 8

[26] Heung-Il Suk, Seong-Whan Lee, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. *NeuroImage*, 101:569–582, 2014. 1

[27] Wei Tan, Prayag Tiwari, Hari Mohan Pandey, Catarina Moreira, and Amit Kumar Jaiswal. Multimodal Medical Image Fusion Algorithm in the Era of Big Data. *Neural Comput & Applic*, July 2020. 1

[28] Atsushi Teramoto, Hiroshi Fujita, Osamu Yamamuro, and Tsuneo Tamaki. Automated detection of pulmonary nodules in pet/ct images: Ensemble false-positive reduction using a convolutional neural network technique. *Medical physics*, 43(6Part1):2821–2827, 2016. 1

[29] Daniel Truhn, Simone Schrading, Christoph Haarburger, Hannah Schneider, Dorit Merhof, and Christiane Kuhl. Radiomic versus Convolutional Neural Networks Analysis for Classification of Contrast-enhancing Lesions at Multiparametric Breast MRI. *Radiology*, 290(2):290–297, Nov. 2018. Publisher: Radiological Society of North America. 1

[30] Robin Wang, Yeyu Cai, Iris K Lee, Rong Hu, Subhanik Purkayastha, Ian Pan, Thomas Yi, Thi My Linh Tran, Shaolei Lu, Tao Liu, et al. Evaluation of a convolutional neural network for ovarian tumor differentiation based on magnetic

resonance imaging. *European radiology*, pages 1–12, 2020. 1

[31] Tao Xu, Han Zhang, Xiaolei Huang, Shaoting Zhang, and Dimitris N Metaxas. Multimodal deep learning for cervical dysplasia diagnosis. In *International conference on medical image computing and computer-assisted intervention*, pages 115–123. Springer, 2016. 1

[32] Adam Yala, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. *Radiology*, 292(1):60–66, May 2019. Publisher: Radiological Society of North America. 8

[33] Youngjin Yoo, Lisa Y. W. Tang, David K. B. Li, Luanne Metz, Shannon Kolind, Anthony L. Traboulsee, and Roger C. Tam. Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from clinically isolated syndrome. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 7(3):250–259, May 2019. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/21681163.2017.1356750. 8