# Evaluating the Role of Multimodal Clinical Data in Breast Cancer Diagnostic Classifier

Si Man Kou, Paul Yomer Ruiz Pinto, Pranshu Datta

**Abstract**—The process of diagnosing breast cancer is fundamentally multimodal in nature. In real-life clinical scenarios, physicians utilise mammography scan images and clinical information to assess cancer. With advances in artificial intelligence, technology is being increasingly integrated into healthcare to support human decision-making. Despite this, several seminal deep learning approaches for breast cancer classification are primarily based on using image or non-image textual clinical data only, without effectively integrating both modalities. Using an established study as a reference, this project compared the diagnostic classification performance of unimodal and multimodal approaches, with the aim of gaining insights into how the performance of a classifier changes when image-level data is combined suitably with non-image clinical data. This project made use of the CBIS-DDSM dataset, which comprises 6,671 breast images from 1,566 participants. Key sampling and preprocessing steps, such as binary encoding and image preprocessing, including pixel value extraction and normalisation, were implemented. Subsequently, the breast images were linked with relevant non-image features, including breast density, mass shape, and calcification type. Unimodal baseline models based on image and non-image data were trained first, following which, the two models with distinct modality fusion strategies were trained. The assessment was conducted using the area under the receiver operating characteristic curve (AUC) and specificity at 95% sensitivity. In the most optimal iteration of the image-only model, an AUC of 0.53 (95% CI: 0.47, 0.6) and a specificity at 95% sensitivity of 10% were achieved. In contrast, the top-performing non-image-only model iteration achieved an AUC of 0.74 (95% CI: 0.67, 0.79) and a specificity at 95% sensitivity of 29%. The finest feature fusion model attained an AUC of 0.67 (95% CI: 0.6, 0.73) with a specificity at 95% sensitivity of 17%. Notably, the learned feature fusion model surpassed the feature fusion model, achieving an AUC of 0.74 (95% CI: 0.68, 0.8) and a specificity at 95% sensitivity of 26%.

**Index Terms**—mammography, image classification, deep learning, transfer learning, multimodal classification, ResNet50

## 1 Introduction

### 1.1 Background and Context

Mammography is a medical imaging technique that utilizes X-rays to examine the human breast for diagnostic and screening purposes. Its primary objective is the early detection of breast cancer, typically achieved through identifying masses or calcifications (Biesheuvel et al., 2011; Braitmaier et al., 2022). However, it's worth noting that the current practice of mammography screening heavily depends on subjective human interpretation for breast cancer diagnosis.

A breast cancer diagnosis is typically multimodal in nature (Holste et al., 2021). This implies that multiple sources of information inform a particular diagnosis. These sources include images obtained from the scans, as well as non-image clinical. Despite this, several seminal algorithms designed for classifying breast cancer diagnoses do not align with the practical application, as they tend to concentrate solely on either image data or non-image data as inputs for their classifiers, without effectively integrating both simultaneously.

The advancements in technology provide an avenue to enhance mammography screening accuracy through the utilization of a larger pool of data via a multimodal approach, thereby mitigating instances of missed cancers and false positives. This is highly important as it underscores the critical significance of achieving more precise breast cancer predictions. Such precision is of paramount value within the medical domain, as it can substantially impact patient outcomes and the overall quality of healthcare. Advanced technologies like machine learning are anticipated to bring significant enhancements to critical aspects of healthcare, including radiology, screenings, and pathology (Ahuja, 2019). These innovations are also expected to offer valuable support to human decision-makers in the breast cancer diagnosis process.

The objective of this project revolves primarily around investigating and analysing the performance of a classifier that utilises multimodal data. Initially, a unimodal methodology is employed, wherein image and non-image clinical data are separately utilized as inputs. This approach facilitates the assessment of the individual effectiveness of each data type and the extent of their contributions to the classification process. Subsequently, this approach is built upon by integrating both the image and non-image data to evaluate the changes in classification performance when adopting a multimodal approach for breast cancer diagnosis.

Through the conducted comparative analysis in this project, insights are sought into the multifaceted characteristics of screening and biomedical data. The primary objective is to elucidate how advanced techniques can be effectively integrated with a diverse array of dynamic clinical data inputs for the accurate classification of breast cancer diagnoses. Ultimately, the project aims to investigate the effect on the performance of a breast cancer diagnosis classifier when integrated multimodal clinical data is supplied as inputs to the classification model.

### 1.2 Literature Review

A systematic literature review was conducted to enhance domain knowledge and comprehension of related work, with a

specific focus on classifiers employing multimodal approaches. Heo et al. (2019) conducted a study that employed chest X-ray imaging data and demographic variables, including age, weight, height, and gender, using CNN models separately. The concatenated features from both models were processed to classify the probability of tuberculosis. This study underscored the significant performance improvement achieved through the inclusion of non-image demographic variables. This study also hypothesized that a larger number of demographic variables might have an even greater positive impact on classifier performance as this study only employed four demographic variables.

Chen et al. (2022) introduced a pathomic fusion approach for the fusion of histology images and genomic data to predict survival outcomes in cancer patients. Their approach leveraged deep learning techniques to model interactions between features from different data modalities. The study demonstrated the effectiveness of this method on glioma and clear cell renal cell carcinoma datasets, highlighting improved prognostic predictions in comparison to using histology or genomics data in isolation.

Spasov et al. (2018) applied CNNs to extract imaging features from MRI data and combined them with structured clinical data, consisting of demographics, genetic information, clinical assessments, and verbal learning data. This integrated data was then used in a feed-forwarded Neural Network for predicting Alzheimer's disease. The study demonstrated the effectiveness of carefully tuning multimodal data, yielding an accurate framework for clinical problems in brain scanning by effectively utilising imaging, continuous, and categorical variables.

While these studies exhibited significant enhancements in classification performance, they often lacked intuitive approaches or overlooked the multiple potential methods for combining information from various modalities. Numerous other research endeavours employed neural networks for addressing distinct problem statements through multimodal data integration (Cheerla & Gevaert, 2019; Huang et al., 2020). However, it's noteworthy that these studies opted for a multi-stage training process instead of pursuing a comprehensive end-to-end training methodology.

In contrast, Holste et al. (2021) expanded upon these findings by implementing an end-to-end training approach. In this research, DCE-MRI data is systematically processed into single-breast images and subsequently integrated with 18 non-image features, encompassing clinical indications and mammographic breast density. The research team has undertaken the training of unimodal baseline models by utilizing image data and clinical data separately. Furthermore, the researchers extended their efforts by developing three multimodal fusion models with the goal of concurrently leveraging both image and clinical data. This allowed the study to assess the best operations to join information from diverse modalities. This research also included a feature importance analysis that aims to discover clinical non-image attributes salient in informing a breast cancer diagnosis within a classifier.

Holste et al. (2021) reported that multimodal fusion models outperformed unimodal models, as reflected in the AUC and specificity at 95% sensitivity metrics. The most successful fusion model attains an AUC of 0.898 and a specificity of 49.1%, highlighting the significant advancements achieved through the integration of non-image data with image data for the purpose of breast cancer classification.

The limitations of this research have functioned as primary drivers for our investigation, necessitating the resolution of these gaps. Principally, the reliance on a particular dataset for classification within this study constrains the potential for broader generalization and the enhancement of performance by the models. Furthermore, the utilization of a dataset that requires imputation of non-image attributes due to missing data necessitates adapting this study to address these constraints.

### 1.3 Motivation for Research

The motivation for assessing the performance of breast cancer classification with multimodal data arises from the acknowledgment that, in the dynamic and evolving field of modern medicine, a breast cancer diagnosis is inherently multifaceted. The incorporation of various data types is directed towards the development of a diagnostic model that closely aligns with the complexities encountered in real-life clinical scenarios. The project aims to explore whether a classification model, supported by diverse multimodal data, can effectively capture the intricacies of the disease, and enhance its diagnostic accuracy. Ultimately, the project endeavours to make a substantial contribution to the field of biomedical data science by investigating the potential advantages of a multimodal approach and facilitating a more comprehensive understanding of the distinct contributions of image and non-image clinical data within the classification process.

## 2 Methodology

In this retrospective analysis, we extend the work applied by Holste et al. (2021), in response to the need identified for generalising their approach over new data. This research aimed to enhance breast cancer diagnosis, considering the complex aspects involved, which aligns with our project's main goals. Furthermore, the study is thorough in its explanation, offering its methodology and sharing accessible source code. Its transparency and reproducibility make it a valuable and flexible resource for future research. Additionally, the Massachusetts Institute of Technology (MIT) license grants free and unrestricted use, modification, and distribution of the software, with no warranty, and absolves the authors from liability for any related claims or damages.

Our main contribution is the examination of the two baseline models and two of the three end-to-end fusion models applied by Holste et al. (2021) in a new dataset. Therefore, to address this challenge we apply image and metadata under sampling, image augmentation and transfer learning. The following research questions have been identified with the aim of gaining a more thorough understanding of the subject:

How does the performance of the two baseline models and two end-to-end fusion models change when using a new dataset compared to the original study by Holste et al. (2021)?

1) *"What are the specific effects of image and metadata under-sampling on theperformance of the baseline models and end-to-end fusion models whenapplied to the new dataset?"*

2) *"How does the incorporation of image augmentation and transfer learningtechniques influence the performance of the two baseline models and two end-to-end fusion models when compared to their performance in the original study?"*

In this section, we outline the process for data preprocessing, modelling and comparing the performance of experiment-varying fusion models, considering factors such as computational demands, available features and the quality of their output (Figure 1).

## 2.1 Data Collection and Description

The Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) dataset was used in this study. The CBIS-DDSM dataset has been improved by radiologists removing incorrectly diagnosed or suspected images from the DDSM dataset, making it suitable for benign and malignant cases (Al-Tam et al., 2022; Smith, 2023). The datasets include cranio-caudal (CC) and mediolateral oblique (MLO) views for both left and right breasts, meaning each patient has four views (Smith, 2023). In our study, we treat each view as a separate image with its associated labels. The CBIS-DDSM dataset comprises 6,671 breast images from 1,566 participants. Essentially, CBIS-DDSM is a modified and standardized version of the original DDSM dataset, focusing exclusively on abnormal cases (benign and malignant), whereas the original DDSM dataset contains 2,620 scanned film mammography images, including normal, benign, and malignant cases (Al-Tam et al., 2022; Smith, 2023). The metadata includes critical information, like mass segmentation, calcification distribution, BI-RADS assessment scores, and pathologic diagnosis, others (Table 1). The CBIS-DDSM dataset is split into a training and testing set based on the abnormality category (i.e., mass or calcification) (Smith, 2023). The split in the dataset is obtained by using an 80%-20% training and testing split (Smith, 2023). The CBIS-DDSM dataset encompasses 753 calcification cases and 891 mass instances (Smith, 2023).

The dataset contains images in DICOM format, providing full scan images, region-of-interest (ROI) images that indicate lesion positions, and cropped images focusing solely on abnormalities without displaying the full mammogram image (Al-Tam et al., 2022; Ansar, 2020; Smith, 2023). For model training and evaluation, we primarily use these cropped breast images as these images provide a clear view of the abnormalities within the breast, excluding extraneous details present in full mammogram images (Figure 2). Meanwhile, ROI images focus simply on locating the areas of abnormalities found in a full mammogram image, and therefore can be challenged to utilise them in cases of breast cancer diagnosis classification (Ansar, 2020; Baccouche et al., 2022).

## 2.2 Data Preparation

After conducting a thorough examination of the datasets provided by the CBIS-DDSM contributors, we selected the metadata based on abnormality type, which was associated with CSV files containing the mass and calcification training and test sets (Lee et al., 2017). Subsequently, we conducted exploratory and correlation analyses of the respective columns. These analyses provided valuable insights from both image and non-image features.

### 2.2.1 Image Approach

An independent analysis revealed the presence of distinct records for cropped images corresponding to each abnormality. This discovery influenced the decision to employ cropped images for our model training. Moreover, it is also inferred from the unique cropped images assigned to each abnormality that the clinical practice of diagnosing breast cancer is potentially based on individual images. According to the dataset, in clinical scenarios, different abnormalities may be identified for different breasts (left or right) or distinct image views (CC or MLO). This dataset variability contributes to the absence of a singular point of reference for consolidating outcomes from various images and mapping them to a single patient context. Therefore, this project adopts a pathology classification approach based on abnormalities present within individual cropped images, rather than patient-based classification. Firstly, each image stored in DICOM format was pre-processed, extracting pixel data from it and storing them as arrays in different lists and dictionaries. Lastly, the cropped images were resized to 224 x 224 pixels converted to a grayscale as a new dimension of the image schema, and its intensity value were linearly normalised to the interval [0, 1] (Holste et al., 2021). These steps produced a final data set of 1438 breast cropped images, each with an associated vector of 33 tabular features and binary diagnosis. These tabular features include clinical and MRI acquisition details and are referred to as "non-image" features.

### 2.2.2 Non-image Approach

Next, continuous non-image features were not standardised as these did not have different scales, and categorical and ordinal features were dichotomized into binary variables via dummy coding. This process resulted in a total of 33 non-image inputs. To ensure a substantial representation of malignant breast cases, we deliberately excluded non-image features that were directly connected to the final breast cancer status from our analysis (i.e., assessment) as they were considered potential false predictor. The inclusion of the 'assessment' variable, which contains BI-RADS information, as a predictor in our study may potentially introduce bias or inaccuracies into our results. This is because BI-RADS serves as a standardized tool that facilitates communication among radiologists and clinicians when categorising and describing breast abnormalities, so its original purpose may directly align with the specific outcome we intend to predict.

The dataset initially consisted of three distinct categories of 'pathology': 'malignant', 'benign' and 'benign_without_callback'. However, a reasonable assumption was made to group 'benign_without_callback' and 'benign' as a single entity (i.e., 'benign'). Moreover, it was also noticed that two nearly identical types of 'calc_type' values, 'lucent_center' and 'lucent_centered' were initially present in the records. These were accordingly rectified to the value 'lucent_center', attributing the presence of two such nearly identical records as a potential result of manual error. Lastly, it was identified that one sample contained a 'breast density' value of '0', which contradicts the established definition of the 'breast density' attribute, designed to have values within a range of 1 to 4. Subsequently, the value was adjusted to reflect a 'breast density' value of 1 for this sample.

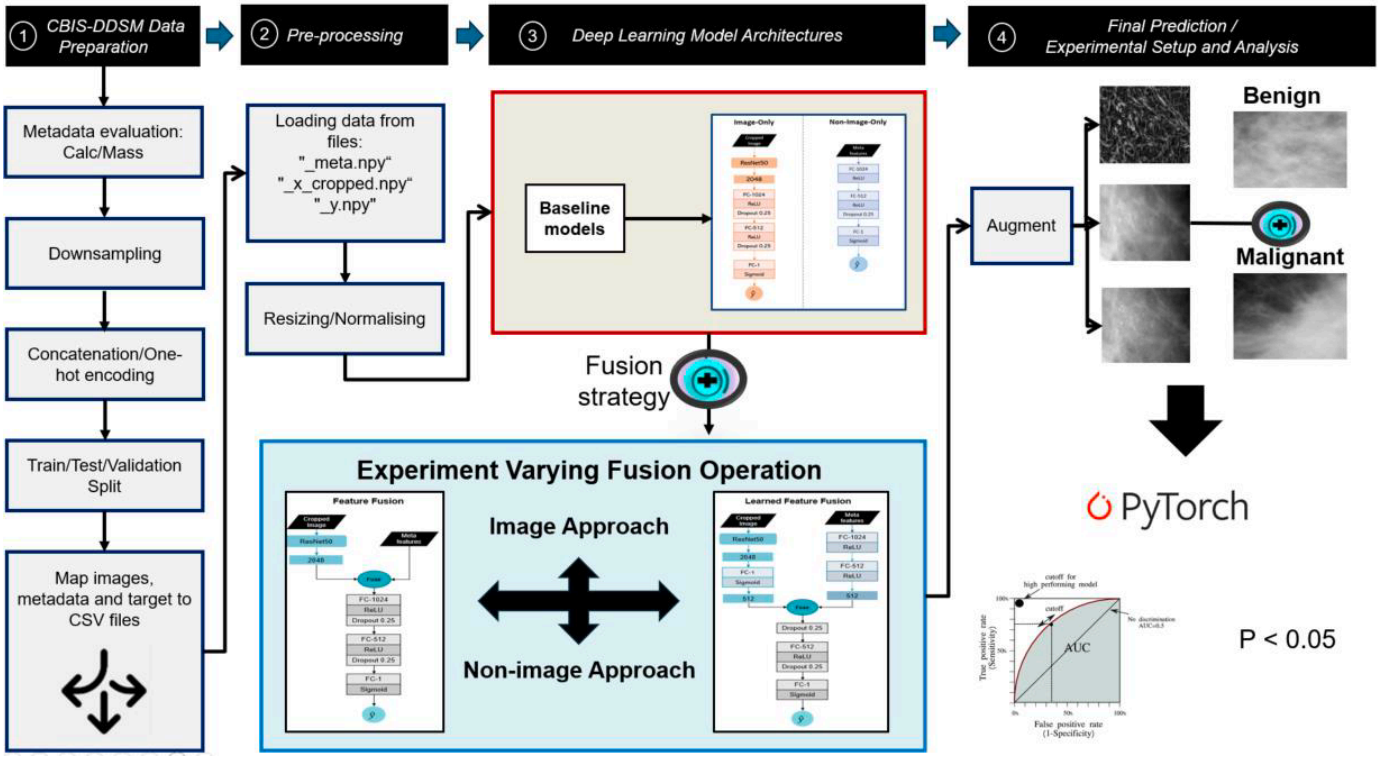


Figure 1: The sequential workflow of the methodology, breaking down each stage, from data preparation to experimental setup and analysis.

| Meta Feature Name | Description |
|---|---|
| Breast density | Category of breast, on a scale of 1 - 4: (1) almost entirely fat, (2) containing scattered areas of fibroglandular density, (3) heterogeneously dense, and (4) extremely dense. |
| Left or right breast | Left or right side of the breast. |
| Image view | CC (craniocaudal) or MLO (mediolateral oblique). |
| Abnormality type | Mass or Calcification. |
| Mass shape | Shape of the mass (when applicable). |
| Mass margin | Mass Margin (when applicable). |
| Calc type | Type of calcification (when applicable). |
| Calc distribution | Distribution type of the calcification (when applicable). |
| Assessment | BI-RADS assessment, on a scale of 0 - 5, describing what was found on a mammogram using 0: need additional imaging, 1: negative, 2: Benign, 3: Probably Benign, 4: Suspicious, 5: Highly suggestive of malignancy. |
| Pathology | Benign, Benign without call-back, or Malignant. |
| Subtlety | Radiologists' rating of difficulty in viewing the abnormality in the image. |

Table 1: Metadata Description.

To ensure uniformity in data and work within the computational limitations, stratified sampling is conducted based on pathology (benign or malignant) for both mass and calcification cases. This stratified approach ensures an equal number of samples for the implementation of experiments while utilising only a portion of the dataset (Table 2). Moreover, following stratified sampling, the mass and calcification records are combined into a single dataset. Categorical columns such as 'left or right breast', 'image view', 'abnormality type', and 'pathology' have been encoded into binary types for programmatic ease and differentiation. For columns with numerous unique records, such as 'mass shape', 'mass margins', 'calc type', and 'calc distribution', one-hot encoding is applied creating new columns (i.e., 'mass_shape_irregular') with binary values. It is also worth noting that some samples contain composite values for calcification type, calcification distribution, mass shape and mass margin. Instead of keeping the composite value as its own single category (i.e., 'calc_type_punctate-pleomorphic'), an additional preparation step was performed to encode the composite value into two separate binary columns (i.e., denoting 'calc_type_punctate' and 'calc_type_pleomorphic' as 1).

Finally, the dataset containing both 'mass' and 'calcification' records, after stratified sampling, is split into training, validation, and test sets in a 6:2:2 ratio, with 80% allocated for training (including validation) and 20% for testing. This

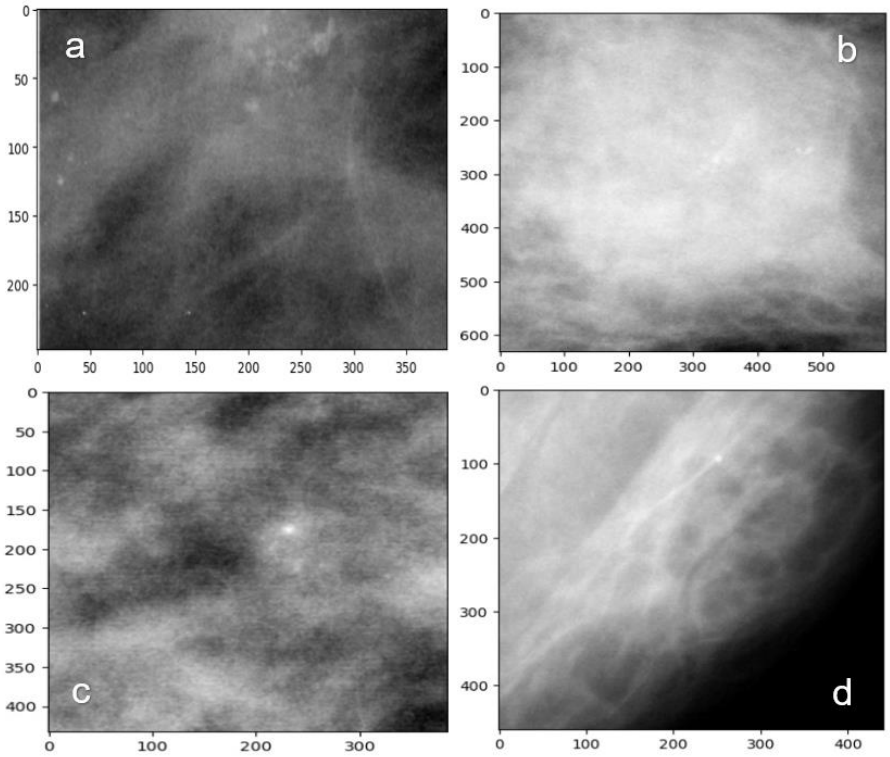


Figure 2: Cropped mammography images, calcification type (a: benign, b: malignant), mass type (c: benign, d: malignant).

split is performed maintaining a balanced 25% sample size split across all combinations of the type of abnormality and the target variable (i.e., calcification benign, calcification malignant, mass benign, mass malignant), as evidenced in Table 3 the data is evenly distributed across calcification and mass abnormality, and benign and malignant across all the three sets.

| dataset | *Before* | | *After* | |
|---|---|---|---|---|
| | Benign | Malignant | Benign | Malignant |
| calc_test | 132 | 127 | 63 | 63 |
| mass_test | 216 | 145 | 72 | 62 |
| calc_train | 607 | 543 | 271 | 271 |
| mass_train | 647 | 626 | 313 | 313 |

Table 2: Stratified sampling across abnormality and pathology - Before vs After. Values represent the distribution of data samples across different datasets before and after stratified sampling. The datasets are categorized by abnormality (calcification and mass type) and pathology (benign and malignant).

### 2.3 Deep Learning Model Architectures

To assess one of the limitations identified by Holste et al. (2021) in their study, if fusion of image and non-image features applied to their models can improve breast cancer prediction in a new dataset, we established the same baseline image-only and non-image feature-only approach. The Image-Only model was built using ResNet50, adapted to accept single-channel input (Holste et al., 2021). ResNet50 is a non-sequential model, composed of a stack of convolutional neural networks (CNNs) combined with residual networks incorporated into each layer; in this architecture, the output of each CNN layer is combined with the original input of that respective layer through residual connections (Seemendra et al., 2021). ResNet50 is widely favoured due to its capacity to manage deep network training, its impressive performance, and the extensive support and resources accessible for its application across diverse computer vision tasks (Seemendra et al., 2021). The Image-Only model was structured with two fully connected layers, and it replaced the initial classification head with a single output neuron. In contrast, the Non-Image-Only model was designed as a straightforward feedforward neural network, consisting of two fully connected layers, and it also featured a single output neuron. (Figure 3).

We explored two of the three primary methods employed by Holste et al. (2021) to combine image-derived features with tabular non-image features. These methods varied in terms of when features from both models were merged in the multimodal architecture. The Feature Fusion model undergoes training to generate a 2,048-feature vector from a breast image. It subsequently combines this vector with the 33 non-image inputs. Through joint learning from both image and non-image features, this combined data is used to make the final prediction (Figure 3). On the other hand, the Learned Feature Fusion model takes a distinct approach by learning features from both the breast image and non-image inputs. It then merges the learned feature vectors from each modality before jointly learning from this amalgamated vector to produce the ultimate prediction. This method allows the model to directly capture interactions between image and non-image features, in contrast to the previous approach, which only combined information at the prediction level (Figure 3).

| | | Training Set | Validation Set | Test Set |
|---|---|---|---|---|
| **Cases** | | 862 | 288 | 288 |
| **Subtlety*** | | 3.69 ± 1.19 | 3.63 ± 1.20 | 3.60 ± 1.21 |
| **Pathology** | Malignant | 431 (50.0%) | 144 (50.0%) | 144 (50.0%) |
| | Benign | 431 (50.0%) | 144 (50.0%) | 144 (50.0%) |
| **Breast Density** | 1 - Almost entirely fatty | 118 (13.7%) | 42 (14.6%) | 46 (16.0%) |
| | 2 - Scattered density | 364 (42.2%) | 111 (38.5%) | 93 (32.3%) |
| | 3 - Heterogeneously dense | 239 (27.7%) | 89 (30.9%) | 102 (35.4%) |
| | 4 - Extremely dense | 141 (16.4%) | 46 (16.0%) | 47 (16.3%) |
| **Breast** | Left | 431 (50.0%) | 153 (53.1%) | 146 (50.7%) |
| | Right | 431 (50.0%) | 135 (46.9%) | 142 (49.3%) |
| **Image View** | MLO | 454 (52.7%) | 157 (54.5%) | 171 (59.4%) |
| | CC | 408 (47.3%) | 131 (45.5%) | 117 (40.6%) |
| **Abnormality Type** | Mass | 462 (53.6%) | 154 (53.5%) | 154 (53.5%) |
| | Calcification | 400 (46.4%) | 134 (46.5%) | 134 (46.5%) |
| **Mass Shape** | Irregular | 152 (17.6%) | 49 (17.0%) | 54 (18.8%) |
| | Lobulated | 120 (13.9%) | 35 (12.2%) | 36 (12.5%) |
| | Oval | 116 (13.5%) | 42 (14.6%) | 36 (12.5%) |
| | Focal Asymmetric Density | 5 (0.6%) | 2 (0.7%) | 1 (0.3%) |
| | Architectural Distortion | 34 (3.9%) | 17 (5.9%) | 13 (4.5%) |
| | Round | 46 (5.3%) | 14 (4.9%) | 18 (6.3%) |
| | Lymph Node | 8 (0.9%) | 5 (1.7%) | 2 (0.7%) |
| | Asymmetric Breast Tissue | 3 (0.3%) | 2 (0.7%) | 2 (0.7%) |
| **Mass Margin** | Spiculated | 129 (15.0%) | 38 (13.2%) | 42 (14.6%) |
| | Ill-defined | 129 (15.0%) | 50 (17.4%) | 50 (17.4%) |
| | Circumscribed | 136 (15.8%) | 47 (16.3%) | 33 (11.5%) |
| | Obscured | 78 (9.0%) | 30 (10.4%) | 31 (10.8%) |
| | Microlobulated | 33 (3.8%) | 14 (4.9%) | 13 (4.5%) |
| **Calc Type** | Pleomorphic | 276 (32.0%) | 91 (31.6%) | 81 (28.1%) |
| | Fine Linear Branching | 46 (5.3%) | 7 (2.4%) | 17 (5.9%) |
| | Amorphous | 48 (5.6%) | 13 (4.5%) | 24 (8.3%) |
| | Punctate | 51 (5.9%) | 29 (10.1%) | 14 (4.9%) |
| | Coarse | 4 (0.5%) | 0 (0.0%) | 1 (0.3%) |
| | Round and Regular | 18 (2.1%) | 8 (2.8%) | 4 (1.4%) |
| | Large Rodlike | 5 (0.6%) | 0 (0.0%) | 2 (0.7%) |
| | Dystrophic | 3 (0.3%) | 2 (0.7%) | 5 (1.7%) |
| | Vascular | 2 (0.2%) | 0 (0.0%) | 3 (1.0%) |
| | Lucent Center | 7 (0.8%) | 5 (1.7%) | 2 (0.7%) |
| **Calc Dist** | Segmental | 64 (7.4%) | 19 (6.6%) | 24 (8.3%) |
| | Regional | 28 (3.2%) | 3 (1.0%) | 9 (3.1%) |
| **Calc Pattern** | Clustered | 269 (31.2%) | 101 (35.1%) | 90 (31.3%) |
| | Linear | 46 (5.3%) | 10 (3.5%) | 15 (5.2%) |
| | Diffusely Scattered | 10 (1.2%) | 6 (2.1%) | 0 (0.0%) |

Table 3: Characteristics across training, validation, and test sets. A "case" comprises a single-breast image and an associated vector of non-image features. Unless specified otherwise, numerical values indicate the count of cases, and values in parentheses indicate the corresponding percentages relative to the total number of cases within a specific set.

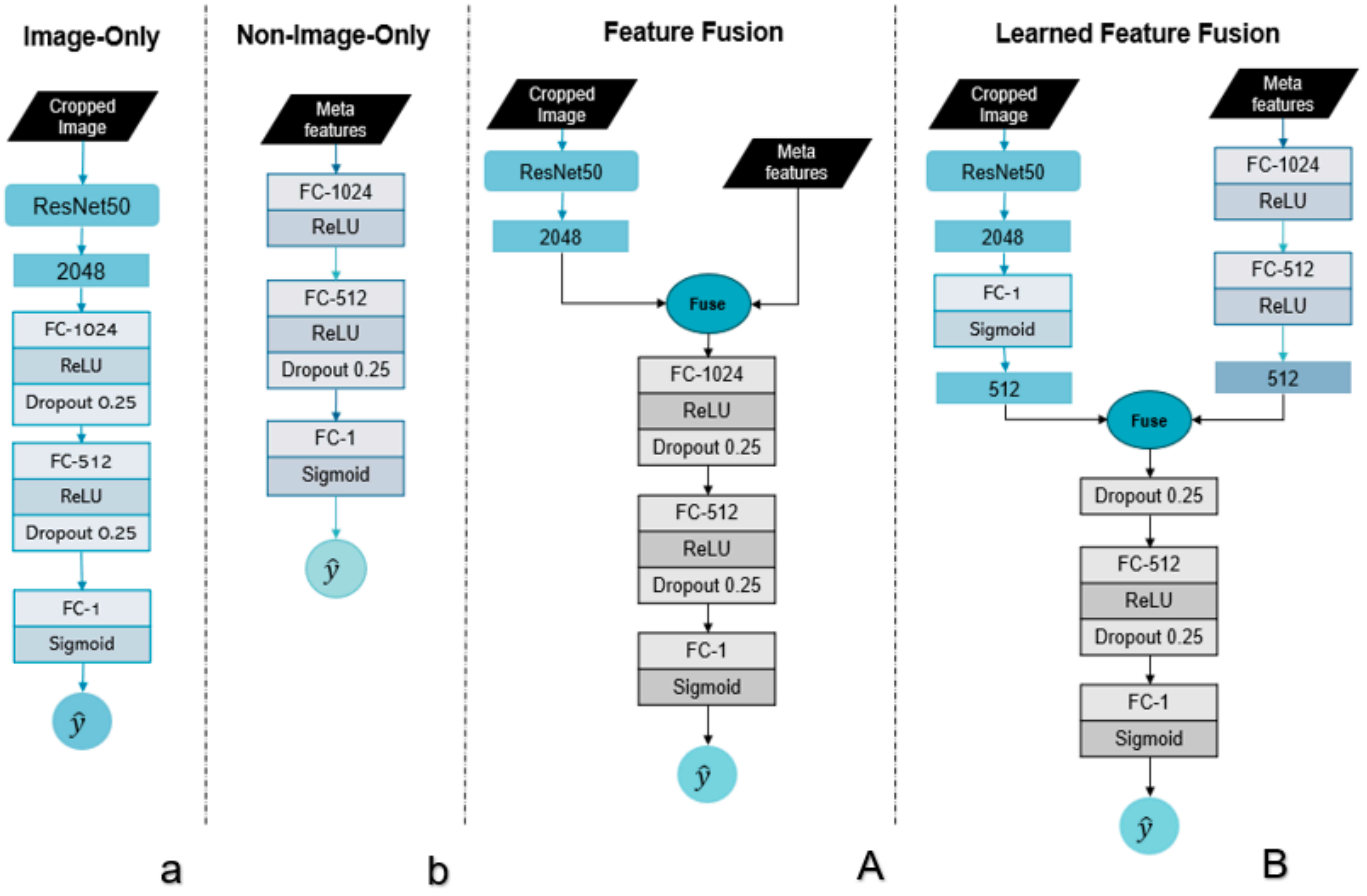


Figure 3: Diagram of fusion architectures that simultaneously incorporate breast imaging and non-image features. Feature Fusion (A) integrates acquired image features with non-image features, while Learned Feature Fusion (B) combines learned image features with learned non-image features. The baseline models, Image-Only (a) and Non-Image-Only (b), are also represented. Feature vectors are depicted without borders, with the number inside denoting the vector's size. "FC-n" designates a fully-connected layer with "n" hidden units, and the symbol "ŷ" signifies the predicted probability

### 2.4 Experiment Varying Fusion Operation

As applied by Holste et al. (2021), concatenation is the primary fusion method to combine information from the models. The architecture adapted and presented in Figure 3 is created to be sufficiently versatile to accommodate alternative fusion operations. To assess the influence of the fusion operation on predictive performance, they also trained different versions of the Learned Feature Fusion model. Instead of concatenating features from each modality, these variants employed elementwise addition (referred to as L+L-Fusion) and multiplication (referred to as L×L-Fusion) on the feature vectors.

While the above models are the same as was intended to assess Holste et al. (2021)'s application over a new dataset, we adapt and modify their approach used for breast MRI fusion data and feature importance analysis due to the differences in the pre-process application of our dataset.

#### *2.4.1 Image Approach*

The breast MRI fusion focuses on handling image data, in our study cropped MRI images. This dataset is designed for training, validation and testing the models mentioned in Section 1.3. that process image data. We apply data augmentation to the images which is a common practice in image-based tasks to improve model generalization (Seemendra et al., 2021). We use "Albumentations", a popular library for image augmentation, to apply various transformations to the images. These transformations include blurring, contrast adjustments, scaling, resizing and using the "sobel filter" (Figure 4). "Sobel filter" is a type of edge detection filter commonly used in image processing which is designed to highlight edges or boundaries within an image (Misra & Wu, 2020). These augmentations introduce diversity in the training data, making the model more robust (Seemendra et al., 2021). The transformed images are converted to "PyTorch" tensors and paired with relevant metadata and labels.

#### *2.4.2 Non-image Approach*

The feature importance analysis focuses on handling non-image features, specifically metadata for feature importance analysis. This dataset is designed for evaluating the impact of metadata features on a pre-trained model's predictions. We do not apply any data augmentation to non-image features. Instead, we ensure that the non-image features and labels are properly converted to PyTorch tensors. The metadata is used for assessing the importance of non-image features using a

pre-trained model. It is a common practice to feed the model with non-image features and analyse how different metadata features influence the models' predictions (Holste et al., 2021).

### 2.5 Experiment Setup and Analysis

We conducted training for all five architectural models as elaborated in Section 2.2.3, in addition to the two variations outlined in Section 2.2.4. The primary objective was to predict breast cancer status. To assess the performance of fusion models, especially in their ability to predict true malignant cases, we compared them to the baseline Image-Only and Non-Image-Only models using two evaluation metrics: (a) AUC and (b)specificity at 95% sensitivity, assessed on the test dataset. Each model underwent training with default values, which implies that no random weight initialization was applied. The training processes for all models remained identical, encompassing the use of the same optimizer, learning rate, data augmentations, and early stopping schedule, among other elements (Table 4). These models were designed and trained utilizing PyTorch version 2.0.1.

We utilised the sklearn.metrics package for computing nonparametric confidence intervals and conducting significance tests to evaluate model performance. All confidence intervals were derived from 10,000 stratified bootstrap samples of the test set, employing the percentile method. To assess the significance of variations in model performance, we applied a straightforward nonparametric test for differences in means. This approach was preferred over the widely-used DeLong test, enabling its use with metrics beyond AUC. A significance level of P-value less than 0.05 was chosen to indicate a statistically meaningful effect.

Lastly, we performed a feature importance analysis for the four models that were trained, as outlined in 2.3 and 2.4. This analysis aimed to evaluate the most influential non-image features for each model. We adopted a permutation-based approach to assess performance, as initially introduced by Holste et al. (2021). This method involved randomly shuffling the values of a single feature in the test dataset, generating new predictions for the permuted data, and subsequently recalculating the AUC. The importance of each feature was determined by the percentage reduction in test AUC following the permutation. The underlying concept is that permuting an important feature should lead to a noticeable change in model performance.

### 2.6 Limitations and Assumptions of the Method

The methodology applied in this study is subject to certain limitations and assumptions. One key limitation lies in the assumption of generalisability from the original study by Holste et al. (2021) to the new dataset. The effectiveness of fusion models that combine image and non-image features may vary due to differences in data pre-processing, data quality, and other dataset-specific factors.

Another limitation pertains to the data augmentation techniques applied to image data. While data augmentation enhances model robustness and generalisation, it also introduces a degree of artificiality into the dataset (Shorten & Khoshgoftaar, 2019). The transformations applied to images may not perfectly reflect the real-world variations in breast abnormalities (Shorten & Khoshgoftaar, 2019). Assumptions are made regarding the pertinence of the chosen augmentation techniques to the task of breast cancer diagnosis, and the impact of these assumptions on the model's performance is not definitively known (Shorten & Khoshgoftaar, 2019).

Furthermore, this study uses image features and non-image features together, and our methodology assumes that these combinations are independent of the specific dataset's characteristics and hold consistent predictive power. However, with non-curated images and metadata, the relevance and influence of non-image features may be sensitive to the dataset's inherent properties, and their impact might vary from one dataset to another (Holste et al., 2021). Careful consideration of these non-image features' suitability for the new dataset is pointed.

Lastly, the assumption of equal significance for the chosen fusion operations (concatenation, addition, and multiplication) should be examined further. While different fusion methods are explored to assess their impact on predictive performance, it is assumed that these operations are equally suitable for the task of combining image and non-image features (Holste et al., 2021). The choice of fusion operation may depend on the specific dataset's characteristics, and its influence on the models' predictions may not be uniform. Therefore, an in-depth analysis of the impact of these fusion operations is necessary to avoid potential biases in model performance.

## 3 Results

### 3.1 Study Cohort

As a result of training, test and validation splits, each split maintained a balanced composition of pathology class and identical distribution of abnormality types, consistently sharing 53.5% of mass cases and 46.5% of calcification cases. The dataset consisted of 862 cropped images for training, 288 for validation, and 288 for testing, with all splits closely mirroring an average subtlety level of approximately 3.6 and a standard deviation of approximately 1.2. Moreover, all three sets share a roughly equal composition in terms of breast density, side of breast, image view, mass shape, mass margin, calcification type and calcification distance, assuring the fact that similar dataset characteristics were used for model learning and validation, thus preventing the risk of overfitting. Scattered and heterogeneously density are the top dominant attributes in terms of breast density. Masses primarily manifested as irregular, lobulated, and oval in shape, while the majority of mass margins were spiculated, ill-defined, and circumscribed. For calcifications, the two most commonly observed profiles are pleomorphic type and clustered distribution.

### 3.2 Model Results

Based on classification results in Table 5, it is notable that image-only is the least effective classified almost akin to a random model, as evidenced by the red ROC curve which closely aligns with the diagonal line, resulting in the lowest AUC value of 0.53 and 0.55 observed respectively in best run and ensemble. Consistent between best run and five run, it also exhibits the highest FNR and lowest sensitivity at 95% specificity, as a result of high number of false predictions on truly malignant cases.

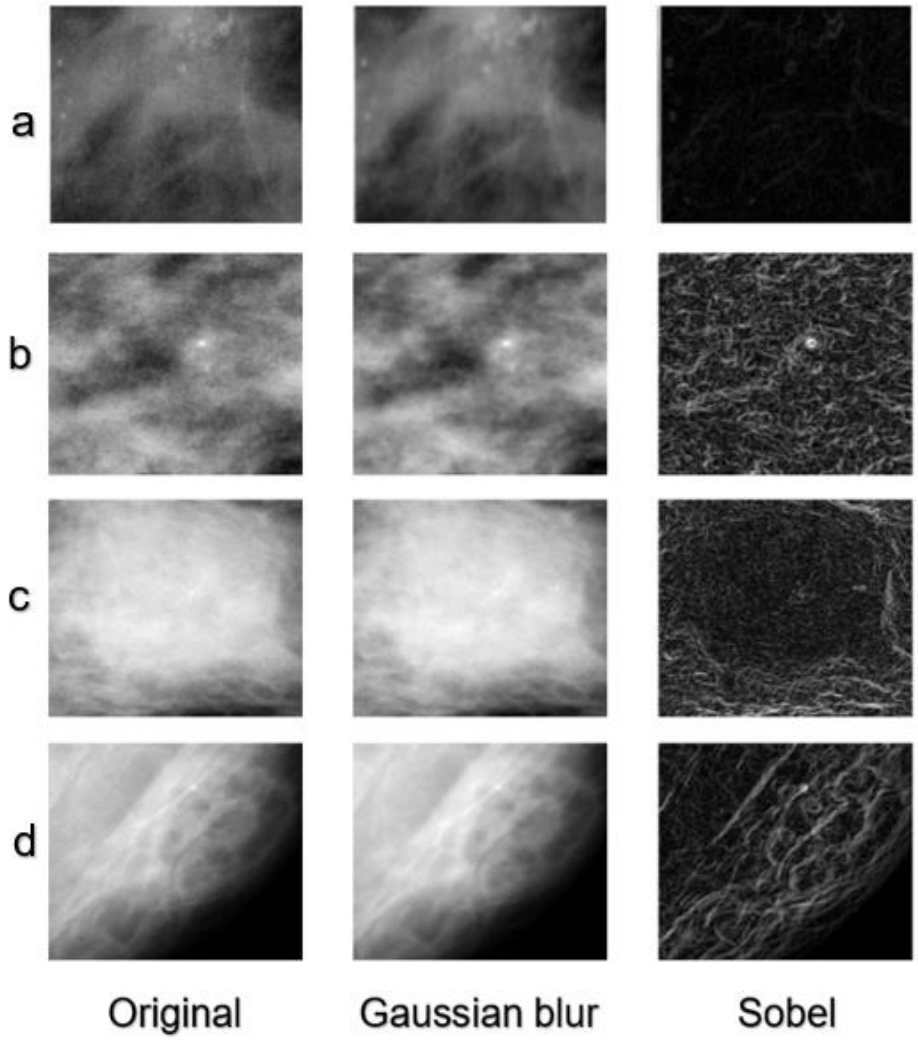


Figure 4: Image comparison of cropped breast image enhanced. Columns represent original pixel value, pixel value after applying Gaussian blur and pixel value of the Gaussian blurred image processed through a Sobel operator. Rows present calcification type (a: malignant, b: benign), mass type (c: malignant, d: benign).

| Model Parameters | Description | Default Value |
|---|---|---|
| **train_mode** | Approach to optimizing fusion model | default |
| **fusion_mode** | fusion type for LearnedFeatureFusion | concat |
| **max_epochs** | maximum number of epochs to train | 100 |
| **batch_size** | batch size for training, validation, and testing | 32 |
| **patience** | early stopping 'patience' during training | 5 |
| **use_class_weights** | whether or not to use class weights applied to loss during training | False |
| **augment** | whether or not to use augmentation during training | True |
| **pretrained** | whether or not to use ImageNet weight initialization for ResNet backbone | False |
| **label_smoothing** | ratio of label smoothing to use during training | 0 |

Table 4: Default Parameters for Model Training

Conversely, non-image-only emerges as the top performing model which achieved the highest sensitivity rates of 29% and 30% for best run and five-run ensemble, along with an ensemble AUC of 0.75. Learned feature fusion exhibits a similar performs but outperforms non-image-only by the best run AUC by 2%, 95% CI by 1% higher lower and upper confidence bounds, and overall FNR by 2-4%. The similar performance between non-image-only and learned feature fusion is also reflected in the interchangeable positions of the blue and yellow ROC curves in Figures 5 despite that learned feature fusion positions itself more closely towards the top-left corner overall in the best model.

Comparing across the two fusion modalities, it is notable that the learned feature fusion outperforms the feature fusion with a significant improvement from image-only across all evaluation measures, suggesting that the joint learning of the learned features of non-image and image is more effective than learning raw non-image features.

### 3.3 Feature Importance

Results in Table 6 shows that mass margin, mass shape and calcification type consistently influenced the most across all models, most dominant in non-image-only with respective values of 3.23, 18.74 and 9.72. It is also noteworthy that non-image-only stands out as the only model with negative importance observed in subtlety, is_left_breast, is_CC_view, is_mass, and calc_dist. Comparing image fusion and non-image-only, a striking reduction in the contributions of these abnormality-related characteristics are evident in the fusion model, as a result of image feature integration.

## 4 Discussion

Utilising a portion of the CBIS-DDSM dataset, we examined the two baseline models and two of the three end-to-end fusion models applied by Holste et al. (2021). We encountered that non-image only model and the multimodal fusion models, which jointly learn from both images and non-images features outperformed image-only model for breast cancer prediction. Notably, the learned feature fusion is the most effective fusion strategy, which means that allowing interaction between learned image-only and non-image features outperformed the approach of combining raw data from an image and non-image features. Furthermore, the non-image only model emerged as best classifier followed by the learned feature fusion. Our findings emphasise that integrating readily available patient imaging and non-image metadata significantly enhances the predictive accuracy of deep learning approaches for automated diagnosis.

The poor performance of the image-only model indicates that using image data alone is not effective for breast cancer prediction. Conversely, the outperformance of the non-image only can be attributed to the intentional limitation of non-image data in our metadata files, which was done for computational purposes. Additionally, the learned feature fusion model demonstrated similar overall performance in comparison to the method applied in the original study (Holste et al., 2021). This suggest that the learned feature fusion model could be considered the ideal model for feature integration and learned representations, making it a promising approach for enhancing the breast cancer detection in the domain.

An analysis of feature importance revealed that attributes such as the mass shape, mass margin, and calcification type held significant predictive power for breast cancer across both non-image-only models and fusion models that incorporated image data. Furthermore, the ranking of relative importance for non-image features exhibited variability among the different models. This variance suggests that specific non-image features may encompass information that can, to some extent, be learnt directly from the images. This observation aligns with the notion that a limited number of non-image features might exhibit a strong correlation with features derived from the images, resulting in their reduced predictive value in fusion models. Additionally, as Holste et al. (2021) pointed out, the variations in feature importance can be attributed to the fact that non-image features do not directly compete with image features within an unimodal model. This could elucidate why the non-image-only model demonstrates higher absolute importance values for nearly every feature in comparison to the learned-feature fusion model. We are presenting these results in an attempt to demonstrate that we observed a similar confounding effect in our data as the one reported in the original study (Holste et al., 2021).

It is recently reported that the use of multimodal learning is anticipated to have a growing significance in precision medicine, serving as a reliable and quantitative method for clinical decision support (Cui et al., 2023). However, multiple studies are still focusing on improving the approach for computer-aided diagnosis based on images-only features or comparing models to identify the optimised deep learning algorithm for identifying breast cancer in mammography pictures (Htay et al., 2018; Kardawi & Sarno, 2023; Ono & Mitani, 2022; Salama et al., 2018; Supriya et al., 2022).

Holste et al. (2021) identified that multimodal approaches might not always be straightforward and could overlook the various ways to integrate information from different modalities. Consequently, they decided to investigate three basic methods for merging features from different modalities and conducted additional experiments to determine the most effective techniques for combining information from these various modalities. Furthermore, they acknowledged a limitation in their models stemming from the retrospective design, which relied on data from a single institution. This highlights the necessity for future research to assess the applicability of their trained models to data from other institutions. In this context, our main contribution is the examination of the two baseline models and two of the three end-to-end fusion models applied by Holste et al. (2021) in a new dataset. When considering their end-to-end fusion methods for breast cancer classification, we find that merging raw data is less effective than combining intermediate learned features from each source. Additionally, we notice that there was no clear difference in performance between combining learned features from both image and non-image sources and using the baseline non-image-only model. This outcome is a result of our deliberate selection of two distinct datasets with a restricted number of attributes, chosen for computational efficiency.

The limitations of this study are primarily due to constraints in computational resources and time, resulting in the use of a relatively small multimodal dataset for training the models. To better understand the impact of expanding non-image features, future research will be necessary to assess

| Model | Best Run | | | | Five-Run Ensemble | | |
|---|---|---|---|---|---|---|---|
| | VAL AUC_ROC | AUC (95% CI) | FNR | Sensitivity at 95% Specificity | AUC (95% CI) | FNR | Sensitivity at 95% Specificity |
| Image-Only | 0.61 | 0.53 [0.47, 0.60] | 0.52 | 0.10 | 0.55 [0.48, 0.61] | 0.39 | 0.09 |
| Non-Image-Only | 0.77 | 0.74 [0.67, 0.79] | 0.28 | 0.29 | 0.75 [0.69, 0.80] | 0.28 | 0.30 |
| Feature Fusion | 0.71 | 0.67 [0.60, 0.73] | 0.22 | 0.17 | 0.64 [0.58, 0.70] | 0.35 | 0.15 |
| Learned Feature Fusion | 0.77 | 0.74 [0.68, 0.80] | 0.26 | 0.26 | 0.74 [0.68, 0.80] | 0.24 | 0.25 |

Table 5: Classification performances of the four models. With respect to AUC with 95% confidence intervals, False Negative Rate and sensitivity at 95% specificity, for both the best run and five-run ensemble. The values for five-run ensemble represent average across five experiment runs trained with different seed numbers, and the run with the highest validation AUC was used to represent the best run. The results for each run were derived from best epoch with the highest early-stopping criterion which is the validation AUC.

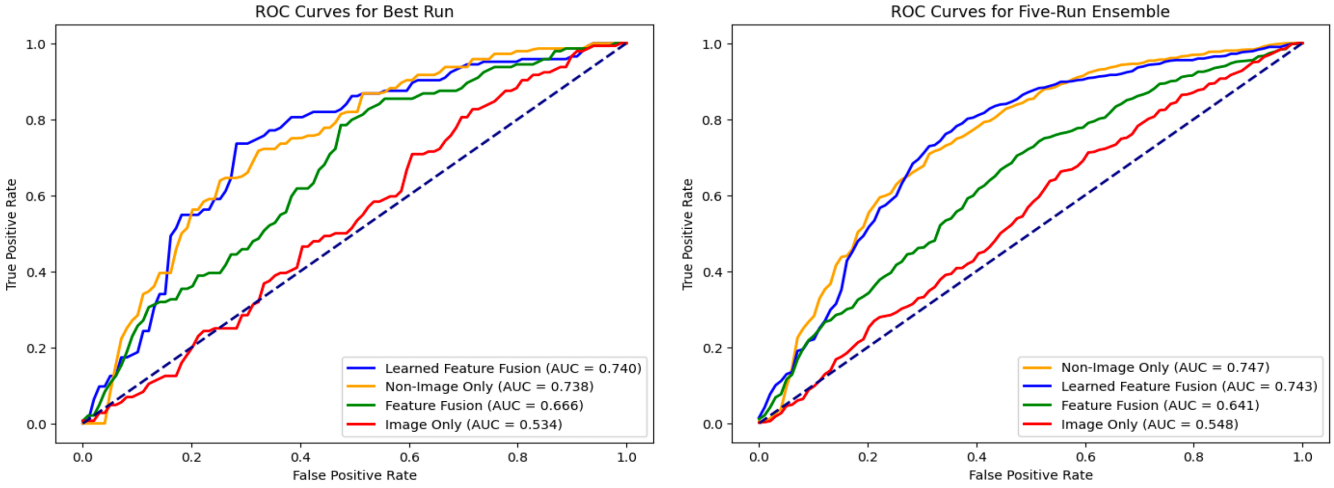


Figure 5: ROC curves and AUC values of the four models respectively for best run (left) and five-run ensemble (right).

| Non-Image Feature | Non-Image-Only | Feature-Fusion | Learned-Feature-Fusion |
|---|---|---|---|
| breast density | 0.84 | 0.02 | 0.20 |
| subtlety | -1.55 | -0.04 | 0.14 |
| is_left_breast | -0.38 | 0.00 | -0.04 |
| is_CC_view | -0.57 | 0.02 | 0.07 |
| is_mass | -1.36 | 0.01 | 0.08 |
| mass_shape | 3.23 | 0.18 | 0.85 |
| mass_margins | 18.74 | 0.14 | 1.24 |
| calc_type | 9.72 | 0.06 | 0.42 |
| calc_dist | -0.71 | 0.01 | 0.12 |

Table 6: Permutation-based feature importance on non-image features. The importance score was calculated by randomly shuffled the values of the associated binary columns once per non-image feature across all the test samples, followed by evaluating the decrease in the new AUC score using the permuted column and the original test AUC. A higher importance value means a larger decrease in observed AUC, suggestive of higher importance in the overall model performance.

whether it can yield similar results in terms of AUC, 95% specificity, and feature importance. Additionally, it's important to note that our non-image features were entirely derived from image features, which could potentially introduce confounding effects into the results.

Regarding the model architectures and training, we aimed to ensure a fair comparison by developing our own data preparation and preprocessing procedures while keeping hyperparameters consistent and not introducing variations in model design. As suggested by Holste et al. (2021), more optimal hyperparameter tuning and architectural choices may lead to slightly different outcomes compared to those presented in this study. Lastly, it's essential to mention that this study utilized cropped images due to memory constraints. Models capable of incorporating full images, region of interest selections from images, and additional sequences from breast MRI exams are likely to further enhance predictive performance.

## 5 Conclusion

The role of multimodal clinical data in the diagnostic classification of breast cancer was evaluated by assessing classification performance separately for image data and non-image attributes. These data sources were subsequently integrated using two fusion strategies: feature fusion and learned feature fusion, which facilitated the integration of multimodal information from the CBIS-DDSM dataset for breast cancer prediction. This project effectively addressed gaps in the initially chosen baseline, enhancing its generalisability and future application in the context of breast cancer classification. Additionally, the project has highlighted the significance of fusion operations in combining data from multiple modalities, underlining that the use of only images or non-image attributes alone is insufficient for accurate predictions. This observation aligns with the inherent multimodal nature of real-life clinical diagnosis scenarios.

In future work, this study can be extended by incorporating a broader range of clinical patient data, including factors such as age, to provide the underlying classifier with more context for diagnostic classifications. Furthermore, enhancements can be made to improve the model's classification performance by introducing penalties for false positive instances where malignant tumours are erroneously classified as benign. Lastly, the inclusion of region of interest (ROI) masking images from the CBIS-DDSM mammography dataset, which was not utilised in this study, could be investigated in future analyses. Such spatial data may offer the classifier a more comprehensive understanding of abnormality distribution across the breast and its implications for diagnosis.

## 6 References


1) Ahuja, A. S. (2019). The impact of artificial intelligence in medicine on the future role of the physician. PeerJ, 7(7702), e7702. https://doi.org/10.7717/peerj.7702
2) Al-Tam, R. M., Al-Hejri, A. M., Narangale, S. M., Samee, N. A., Mahmoud, N. F., Al-Masni, M. A., & Al-Antari, M. A. (2022). A Hybrid Workflow of Residual Convolutional Transformer Encoder for Breast Cancer Classification Using Digital X-ray Mammograms. Biomedicines, 10(11). https://doi.org/10.3390/biomedicines10112971
3) Ansar, W. S., A.R.; Raza, B.; Dar, A.H. (2020). Breast Cancer Detection and Localization Using MobileNet Based Transfer Learning for Mammograms Intelligent Computing Systems. ISICS 2020
4) Baccouche, A., Garcia-Zapirain, B., & Elmaghraby, A. S. (2022). An integrated framework for breast mass classification and diagnosis using stacked ensemble of residual neural networks. Sci Rep, 12(1), 12259. https://doi.org/10.1038/s41598-022-15632-6
5) Biesheuvel, C., Weigel, S., & Heindel, W. (2011). Mammography Screening: Evidence, History and Current Practice in Germany and Other European Countries. Breast Care (Basel), 6(2), 104-109. https://doi.org/10.1159/000327493
6) Bradford, J., & Perrin, D. (2019). A benchmark of computational CRISPR-Cas9 guide design methods. PLoS Comput Biol, 15(8), e1007274. https://doi.org/10.1371/journal.pcbi.1007274
7) Braitmaier, M., Kollhorst, B., Heinig, M., Langner, I., Czwikla, J., Heinze, F., Buschmann, L., Minnerup, H., Garcia-Albeniz, X., Hense, H. W., Karch, A., Zeeb, H., Haug, U., & Didelez, V. (2022). Effectiveness of Mammography Screening on Breast Cancer Mortality - A Study Protocol for Emulation of Target Trials Using German Health Claims Data. Clin Epidemiol, 14, 1293-1303. https://doi.org/10.2147/CLEP.S376107
8) Cheerla, A., & Gevaert, O. (2019). Deep learning with multimodal representation for pancancer prognosis prediction. Bioinformatics, 35(14), i446-i454. https://doi.org/10.1093/bioinformatics/btz342
9) Chen, R. J., Lu, M. Y., Wang, J., Williamson, D. F. K., Rodig, S. J., Lindeman, N. I., & Mahmood, F. (2022). Pathomic Fusion: An Integrated Framework for Fusing Histopathology and Genomic Features for Cancer Diagnosis and Prognosis. IEEE Trans Med Imaging, 41(4), 757-770. https://doi.org/10.1109/TMI.2020.3021387
10) Cui, C., Yang, H., Wang, Y., Zhao, S., Asad, Z., Coburn, L. A., Wilson, K. T., Landman, B. A., & Huo, Y. (2023). Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. Prog Biomed Eng (Bristol), 5(2). https://doi.org/10.1088/2516-1091/acc2fe
11) Heo, S. J., Kim, Y., Yun, S., Lim, S. S., Kim, J., Nam, C. M., Park, E. C., Jung, I., & Yoon, J. H. (2019). 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). https://doi.org/10.3390/ijerph16020250

12) Holste, G., Partridge, S. C., Rahbar, H., Biswas, D., Lee, C. I., & Alessio, A. M. (2021). End-to-End Learning of Fused Image and Non-Image Features for Improved Breast Cancer Classification from MRI 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW),
13) Htay, T. T., Yangon, M., Maung, S. S., & Yangon, M. (2018). Early Stage Breast Cancer Detection System using GLCM feature extraction and K-Nearest Neighbor (k-NN) on Mammography image The 18th International Symposium on Communications and Information Technologies (ISCIT 2018), Bangkok, Thailand.
14) Huang, S. C., Pareek, A., Zamanian, R., Banerjee, I., & Lungren, M. P. (2020). Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. Sci Rep, 10(1), 22147. https://doi.org/10.1038/s41598-020-78888-w
15) Kardawi, M. Y., & Sarno, R. (2023). Image Enhancement for Breast Cancer Detection on Screening Mammography Using Deep Learning 2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE),
16) Lee, R. S., Gimenez, F., Hoogi, A., Miyake, K. K., Gorovoy, M., & Rubin, D. L. (2017). A curated mammography data set for use in computer-aided detection and diagnosis research. Sci Data, 4, 170177. https://doi.org/10.1038/sdata.2017.177
17) Misra, S., & Wu, Y. (2020). Machine learning assisted segmentation of scanning electron microscopy images of organic-rich shales with feature extraction and feature ranking. In Machine Learning for Subsurface Characterization (pp. 289-314). https://doi.org/10.1016/b978-0-12-817736-5.00010-7
18) Ono, Y., & Mitani, Y. (2022). Evaluation of feature extraction methods with ensemble learning for breast cancer classification 2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech),
19) Salama, M. S., Eltrass, A. S., & Elkamchouchi, H. M. (2018). An Improved Approach for Computer-Aided Diagnosis of Breast Cancer in Digital Mammography 2018 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Rome, Italy.
20) Seemendra, A., Singh, R., & Singh, S. (2021). Breast Cancer Classification Using Transfer Learning Evolving Technologies for Computing, Communication and Smart World. Lecture Notes in Electrical Engineering, Singapore.
21) Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. Journal of Big Data, 6(1). https://doi.org/10.1186/s40537-019-0197-0
22) Smith, K. (2023). Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) https://wiki.cancerimagingarchive.net
23) Spasov, S., Passamonti, L., Duggento, A., Lio, P., & Toschi, N. (2018). A Multi-modal Convolutional Neural Network Framework for the Prediction of Alzheimer Disease Annu Int Conf IEEE Eng Med Biol Soc,
24) Supriya, U., Madhumathi, R., & Sulthana, R. (2022). An Analysis of Deep Learning Models for Breast Cancer Mammography Image Classification 2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI),