Finding Hidden Features Responsible For Machine Learning Failures In Breast Cancer Detection on Screening Mammography

Frederik Bechmann Faarup
frfa@itu.dk

Supervisor
Veronika Cheplygina
vech@itu.dk

Abstract

With promising results, machine learning models can help early detection of breast cancer and thereby increase the odds of survival. However, clinical applications may still be limited, as model decisions could be misguided by hidden features, such as text overlay, radiologist markers and data set characteristics. Through error analysis of existing models, I show that performance differs on mammography screenings with certain hidden features present. Further, correlations between hidden features and model posterior probabilities were found. More work on bigger data sets is warranted to truly conclude the effects of hidden features on model outputs. Transfer learning showed promising results in combating data set variance as a hidden feature, and it enables further investigations on new data sets, potentially revealing the effects of different hidden feature types.[1]

1 Introduction

Breast cancer accounts for 23.3% of all cancer cases for women (Kræftens Bekæmpelse, 2020), and it is the second-largest cause of cancer-related deaths in developed countries (Bray et al., 2018). Early detection plays a crucial role in increasing the odds of survival (Cancer.org, 2022), but it requires the work of professional radiologists. Machine learning can be used as a tool to reduce the workload of radiologists, allowing more patients to be screened (Shoshan et al., 2022). Before clinical application, decision-making of the models should be investigated, as they may be biased by hidden features in the data.

Hidden features are in other works (Budrys et al., 2018) described as ”artifacts”, but the more general term *hidden features* also encapsulates data set variance, inconsistent background coloring, post-process markers etc.

[1]Github: https://github.com/frfa1/bsc

This report investigates how machine learning models, particularly CNNs based on Resnet and VGG, are affected by hidden features present in mammography screenings. The models from research are examined on the official CBIS-DDSM test set with 645 whole mammograms, and on a stratified test set of INbreast, both public data sets.

2 Background Terms

Mammogram: A mammogram is an X-ray image of a breast. It is used as a tool to detect and diagnose breast cancer at early stages in what is referred to as a mammography screening.
MLO and CC views: Bilateral craniocaudal (CC) and mediolateral oblique (MLO) are two types of views used in most mammography screening routines, hence they are part of the *standard views* group. They both depict the breast but from different angles.
Mass and Calcification: Mass and calcification are two different types of abnormalities that can be indicative of cancer. Such abnormalities can be either benign (normal) or malignant.
DICOM: Digital Imaging and Communications in Medicine (DICOM) is a special image format used in medical imaging.
ROI: Region of interest (ROI) is used in this project as a segmented area of a medical image, either done by radiologists or by a segmentation model. ROIs are used to delimit lesions or other abnormalities on mammograms.
BI-RADs: Breast Imaging-Reporting And Data System (BI-RADS) is a widely used tool for assessing and reporting risk in breast imaging, and thereby mammography. BI-RADS consists of the following seven categories, which breast images are assessed to upon screening by a radiologist:

0: Incomplete and additional imaging is required before evaluation.

1: Negative (i.e. no findings, not even a benign finding).

2: Benign.

3: Probably benign.

4: Suspicious of malignancy.

5: High probability of malignancy.

6: Biopsy-proven malignancy.

3 Related Work

In the field of screening mammography, much work has been done to achieve results similar to or better than radiologists (McKinney et al., 2020; Henn et al., 2021) with the use of machine learning models in non-clinical environments. While studies have shown clinical utility of the current state of machine learning in mammography (Batchu et al., 2021), there still seems to be implications and pitfalls to be handled.

For example, although cutting false-positives by up to 5.7% and false-negatives by up to 9.4%, a large study by Google (McKinney et al., 2020) has been criticized for ignoring racial differences in mammography. As seen in other studies (Wu et al., 2019a), a racial biases could be perpetuated into the core of the machine learning models presented in the research. Implementing such a model in a real-world mammography screening setup could have negative consequences for minority groups, or other groups affected by other biases in the models.

Other studies show, that while achieving high AUROC scores on their own validation sets, several deep learning models in mammography research have low levels of robustness, when tested on an external validation set (Wang et al., 2020). This indicates that the models do not generalize well, and it opens up the possibility of biases in the training data at different hospitals, originating from different scanner types, postprocessing, sampling methods etc.

In related fields, robustness has been shown unclear, too, as models rely on hidden features rather than medical pathology. This was shown in radiographic COVID-19 detection (DeGrave et al., 2021), where common approaches used to collect data enabled the models trained on the data to take false shortcuts. Through explainable AI and saliency maps, it was found that swapping laterality markers on the input images changed the model outputs significantly. In other words, a subtle, hidden feature had affected the model outputs.

Motivated by this, I investigate the effects of hidden features on machine learning models in breast cancer detection. I challenge model performance on data sets commonly used in research, and evaluate how the models are affected by the hidden features present on mammograms.

4 Data

4.1 CBIS-DDSM

CBIS-DDSM (Lee et al., 2017) is one of the most known public mammography data sets consisting of full mammograms, ROI segmentations (extracted by radiologists), and a set of image- and ROI-level features. The ROI segmentations are part of the reason why this particular data set is widely used, as it is an expensive and time consuming task to generate and label pixel-level ROIs. Table 1 displays the features of CBIS-DDSM from the meta data files, where redundant columns (e.g. all empty or single-valued columns etc.) are omitted.

Column	Description
patient_id	A unique ID for the patient
breast_density	The density of the breast
left or right breast	Whether the image is left or right
image view	The view, either CC or MLO
assessment	BI-RADS assessment
pathology	The pathology; benign, benign without callback or malignant
subtlety	Mammographer-assigned ”subtlety rating” (1 to 10)
abnormality type	The type or abnormality, either mass or calcification
...	...
mass shape	The shape of the mass
mass margins	The margins of the mass
calc type	The type of calcification
calc distribution	The distribution of the calcification

Table 1: Mass/Calcification Case Description Sets for both train and test in CBIS-DDSM. The last four columns corresponds to either mass or calcification.

The CBIS-DDSM data set has a standardized train/test split, where the test set consists of a vari-

ation of cases in terms of difficulty, i.e. the data is split in such a way that both sets have the same distribution of BI-RADS. This split was made separately for mass cases and for calcification cases, as a way to ensure a balanced stratification.

The total data set contains 1.566 patients with 3.103 full mammogram images. Out of this, the official test set consists of 349 patients with 645 full mammogram images (split on different views and laterality). Each of those 645 images can have several ROIs with the same or different verified pathology label. The 349 patients are further split into 151 calcification cases and 201 mass cases - exceeding 349 because some patients are on a ROI-level counted in both case types. The full mammogram images contain ROI-level BIRADs assessments, too, which are assessed by radiologists, and are not the final, true pathology. Figure 1 shows the distribution of BI-RADS in the official CBIS-DDSM test set, which will be useful later, when models are evaluated.

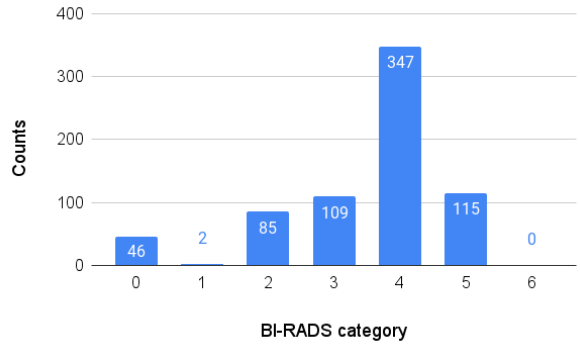


Figure 1: BI-RADS distribution of the official CBIS-DDSM test set.

All the images from the CBIS-DDSM data set are derived from multiple different scanners at different institutions. The data set is used in this project because of its availability, extensive documentation, variation of cases and because it is standardized, allowing a fair comparison between different, existing models. Lastly, and perhaps most importantly, it contains cases with visible, hidden features that should not - but could - affect machine learning models trained on the data.

4.2 INbreast

Another public mammography data set is INbreast (Moreira et al., 2012), which originates from a breast center in a university hospital in Porto (*Centro Hospitalar de S. Joao [CHSJ]*). Structurally, it

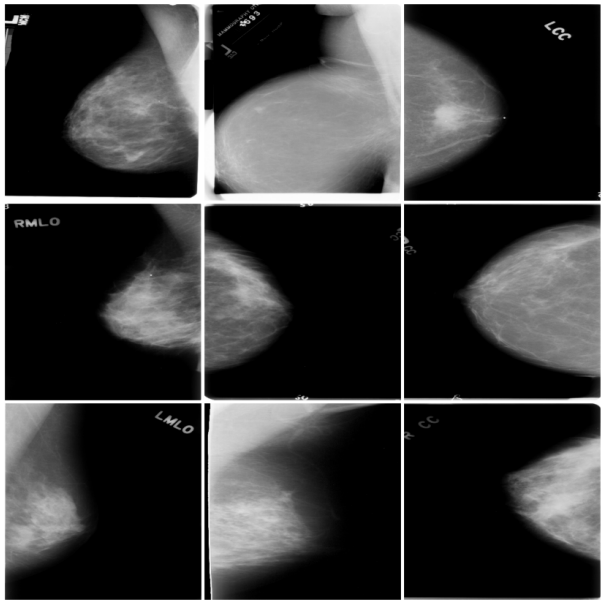


Figure 2: Examples of full mammograms from CBIS-DDSM.

contains a folder with all the full mammogram DICOMs, a folder with ROI segmentations, a folder with XML-files describing the ROIs, a folder containing TXT-files with patient medical history in Portuguese, and lastly, a metadata CSV-file containing the information shown in table 2.

The data set consists of 115 patients spread on 410 full mammogram images. It contains a variety of type of cases, i.e. a variety of BI-RADS cases and cases with different types of abnormalities (mass, calcification, asymmetries and distortions). The distribution of BI-RADS in the whole data set can be seen in figure 3.

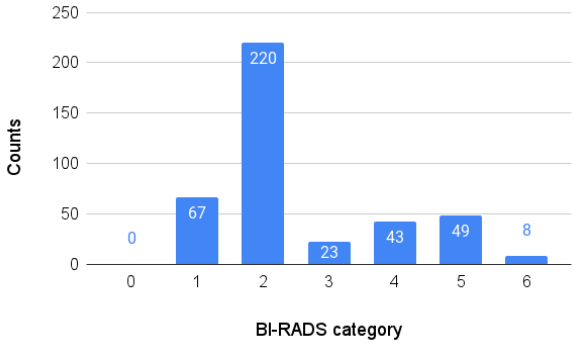


Figure 3: BI-RADS distribution of INbreast.

Likewise to CBIS-DDSM, the INbreast data set is somewhat well-documented, and it has a good stratification of cases in terms of BI-RADS. What makes INbreast particularly interesting in

Column	Description
Laterality	Which side the breast is, L or R.
View	The standard view, CC or MLO.
Acquisition date	The acquisition date of the mammogram.
File Name	The filename.
ACR	A measure of density in ACR standard scale.
BI-RADS	Radiologist assessed BI-RADS
Mass / Micros / Distortion / Assymetry	Whether the image contains the feature. X or None for each feature.
Findings Notes (in Portuguese)	One or a couple of keywords in Portuguese describing the mammogram.

Table 2: General metadata-file for INbreast.

this project, is its visible differences compared to CBIS-DDSM, as seen when comparing the previous figure 2 and figure 4. This difference in visual appearance (and therefore also data set characteristics) stems from the fact that INbreast is built with full-field digital mammograms, while CBIS-DDSM is built with scanned and digitized film mammograms. Further, they originate from different scanners at different hospitals.

5 Methodology

5.1 Preprocessing

The following section describes, how the data is preprocessed. Starting off with **CBIS-DDSM**, it was downloaded from the The Cancer Imaging Archive (TCIA) website. All the DICOM images were converted to 16-bit PNG-format. Each DICOM image was read as a pixel array with the Python *pydicom* library, and then the *PyPNG* library was used to create a new greyscale PNG image, where the content of the pixel array was stored. Later, before model evaluation, each image was downsized to a height of 1152 and width of 896 in pixels, and rescaled to maximum pixel value of 255.

The meta data for the CBIS-DDSM test set contained more rows than there were full mammogram images, because some mammograms contained multiple ROIs. This resulted in some of the mammograms containing ambiguous features and labels - for example, if a mammogram contained both a benign and a malignant lesion. The ambiguous labels were handled by labeling the full mammograms in the following, mutually exclusive way;

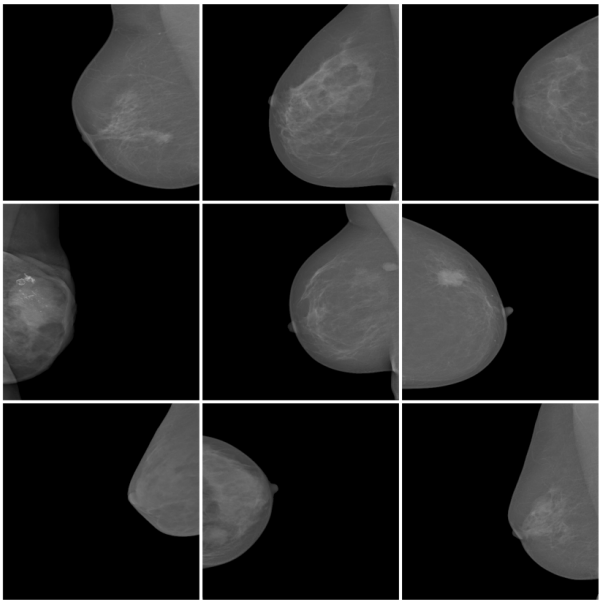


Figure 4: Examples of full mammograms from INbreast.

- Malignant if any malignant ROI present
- Benign otherwise

Moving on to **INbreast**, it was preprocessed similarly to CBIS-DDSM, by wrangling the meta data and converting the full mammogram DICOM files to 16-bit PNG format. The data set did not contain verified pathology for each full mammogram. For that reason, the same labeling method was chosen as in (Wu et al., 2019b):

- Malignant if BI-RADS 4, 5 or 6
- Benign if BI-RADS 1 or 2
- Discarded if BI-RADS 3, because the assessment is typically not given at screening (12 patients with 23 images).

As the data set did not consist of an official train/test split, a split was made as part of the preprocessing. 60% of the total set was used in train, 20% in validation and 20% in test. This differs from the end2end paper, as they omit the test set

and opt for a 70-30 train/validation split instead. The paper reports a high AUROC on their validation set, but as model hyperparameters are fine-tuned on the validation set, it may not generalise well on an unseen test set. As described in (Wang et al., 2020), models in mammography machine learning research tend to perform a lot better on validation sets over test sets. For that reason, I chose to leave out a final, unseen test set to evaluate on.

The split was made in a grouped, stratified fashion, so that all images for one patient was in the same set, and so that each set had a similar distribution of BI-RADS. This was done to ensure that the test set would not evaluate patients, which it had seen during training, and so that the difficulty of the sets were similar.

5.2 Identifying Potential Hidden Features

Several sources were found to outline the possible hidden features, that commonly occur in mammograms (Ayyala et al., 2008; Chaloeykitti et al., 2006). Meanwhile, the whole set of CBIS-DDSM test was manually looked through to find extra hidden features that were not described in literature. This combined process resulted in a list of possible hidden features. After making the list, each image in the CBIS-DDSM test set was thoroughly looked through to check if each possible feature was present on the mammogram. If a given image contained a certain feature, then it was noted in a meta data file (available on the Github). Table 3 outlines each possible hidden feature and the number of times it was present in the CBIS-DDSM test set.

Figures 5(a-d) exemplify a set of the hidden features to give the reader a visual understanding of their appearance. Ideally, a radiologist would confirm the hidden feature labels, as I by no means have the medical experience to do such - especially in a setting, where the results would have a real world application. While some of the features, like *text*, were relatively easy to spot in the images, others were more subtle. For example, the *nipple marker* feature, which is drawn as a perfectly round circle of brightness, resembled birthmarks with a round shape and a similar size. For that particular feature, the image was labeled only if the bright, round object was *perfectly* round, and otherwise not. In cases with doubt, the image was not labeled. This restrictive labeling practice was chosen as a way to avoid over-labeling a feature that could be an actual, true indication of breast cancer, like a type of calcification. Figure 5(d) compares an image of a round birthmark with a round nipple marker.

Name	**Hidden Feature**	**Counts**
F1	Text	471
F2	An unidentifiable object looking like a ruler	97
F3	Nipple marker	86
F4	Scar marker	18
None	A pacemaker	0
None	An implant	0
None	Mole marker	0
None	Palpable mass marker	0
None	Area of concern marker	0

Table 3: Counts of full mammograms with each hidden feature present in the CBIS-DDSM test set.

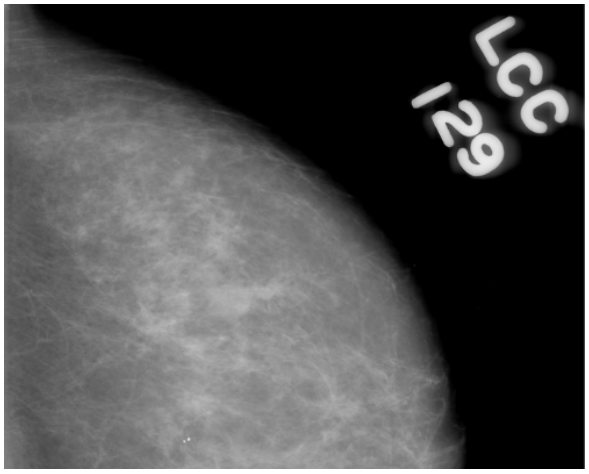
Many images had a perfectly round circle, resembling the nipple marker, on locations away from the nipple, sometimes even outside of the breast. I speculate that this is either because the digitization process has reshaped the image without changing an overlay with the marker similarly, or because the marker symbol has an ambiguous meaning at the different local hospitals from where CBIS-DDSM originates. Nonetheless, all these type of cases were labeled.

5.3 Choice of Model

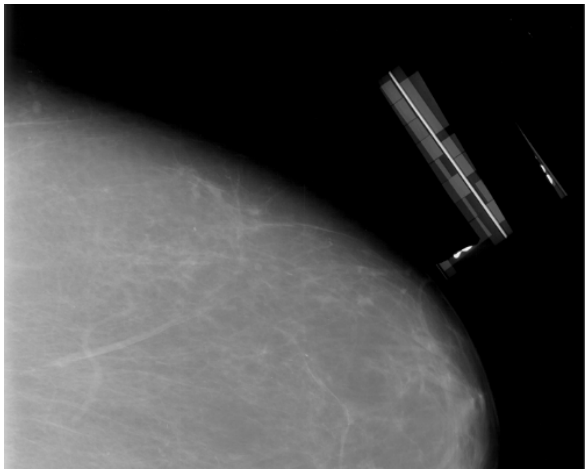
As this project sought to evaluate the effects of hidden features on machine learning models, it made sense to find existing, well-performing models in the field. Two deep learning models were considered, due to their code availability, published results and design-choices:

1. **(Wu et al., 2019b)**
2. **(Shen et al., 2019)**

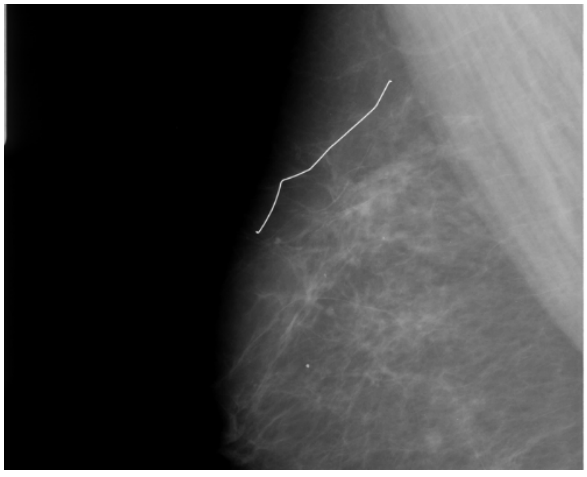
The former model was initially tested on the CBIS-DDSM test set, but near random results were achieved. As the model data was private, and as



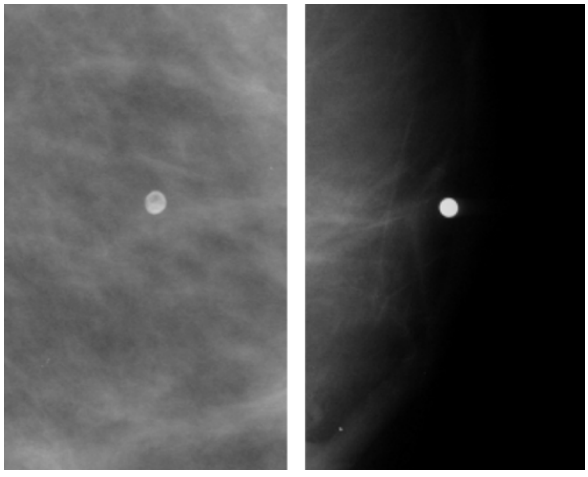
(a) An example of a mammogram with text present.



(b) An example of a mammogram with an unidentifiable "ruler" object present.



(c) An example of a mammogram with a scar marker present.



(d) Left: An example of a round birthmark. Right: An example of a nipple marker.

Figure 5: Examples of hidden features in the CBIS-DDSM test set.

better-than-random results were needed to be able to isolate the effects of the hidden features, this particular model was discarded.

The latter was trained on the CBIS-DDSM train set, and the documentation allowed experiments to be set up from already trained models. As described in the following section, a particular useful aspect of the model by (Shen et al., 2019) is its two-step approach, which firstly trains a patch classier, and then on top of that trains a whole image classifier. This type of complex approach is referred to as an *end2end model*, which is also what the models from the paper are referred to throughout this report.

5.4 Model Architecture

The end2end model is built up of two components, that make up the whole image classifier:

1. **A patch classifier**, based on pre-trained Resnet50 or VGG16.

2. **Classification layers**, containing blocks of Resnet or VGG on top of the patch classifier layers.

This approach was chosen by (Shen et al., 2019) because ROI-level labels are generally rare in medical imaging, as they are expensive and time consuming to gather. They proposed models that pre-train a patch classifier on available data with ROI-level labels, and then further transfer learn a whole image classifier on more available data with image-level labels.

The patch classifier is trained on patches extracted from the whole images in the CBIS-DDSM train set. Many patches were sampled for each image in the data set, both around ROIs and from the backgrounds. The patches were labeled using the true ROI labels (benign or malignant), or as "background". The labels were also divided into mass or calcification cases, resulting in a total of five classes: *background, malignant mass, benign mass, malignant calcification* and *benign calcification*. To compare the performance, the authors trained patch classifier models from scratch and pre-trained on ImageNet. As the latter performed significantly better, those are the models used in this project.

The patch classifier is a convolutional neural network based off of either the 16-layer VGG network (Simonyan and Zisserman, 2014) or the 50-layer residual network (Resnet) (He et al., 2015). In short, the VGG network uses 3x3 filter sizes with a stride of 1 and 2x2 max pooling. The smaller filter sizes reduce the number of trained parameters substantially, while the network is still able to achieve similar feature maps as convolutional networks with larger filter sizes. The Resnet blocks on the other hand use a stride of 2 instead of 2x2 max pooling in the first convolutional layer, which immediately reduces the feature map size. Then, it is followed by multiple convolutional layers. The unique part about Resnet are the residual blocks, which push layer outputs through a shortcut and onto the 2nd or 3rd consecutive layer. Empirically, this feature allows Resnets to go deeper with the layers, without performing worse than their shallower counterparts.

When the patch classifier is used singly to classify patches, it applies the Softmax activation function to output the probabilities per class on a given image patch input:

$$f(z)_j = \frac{e^{z_j}}{\sum_{i=1}^{c} e^{z_i}} \text{ for } j = 1, ..., c \qquad (1)$$

By sweeping over whole mammograms with the patch classifier, a grid of probabilities, alias a *heatmap*, can be produced, which is utilized later on.

On the other hand, when the patch classifier is used as an intermediate step in the whole image classifier, the ReLu activation function can be used instead. This is done as a way to avoid diminishing gradients for large inputs, which is associated with Softmax:

$$f(z)_j = max(0, z_j) \text{ for } j = 1, ..., c \qquad (2)$$

However, the authors of the models achieved the best performance when they removed the heatmap altogether in the whole image classifier, and instead added the top layers directly onto the final layers of the patch classifier. That particular setup is the one used as whole image classifiers throughout this project.

The top layers consist of two convolutional layers, again in the form of either VGG or Resnet, followed by a global average pooling layer, and then the final image classification output. A hybrid model is also considered in this project, in which a patch classifier based on VGG uses two Resnet layers as top layers. With the pure VGG/Resnet

models and the Hybrid model, this makes up a combined total of three whole image classifiers, that were evaluated in this project.

6 Experiments

6.1 Experimental setup: CBIS-DDSM

Three experiments that sought to evaluate the effect of hidden features were conducted on the CBIS-DDSM data set:

1. Testing pre-trained models on the CBIS-DDSM test with different hidden feature subsets removed
2. Sweeping a patch classifier over false-positive and false-negative cases of each feature to generate example heatmaps
3. Multiple linear regression on CBIS-DDSM test to evaluate the correlation between each hidden feature and model posterior probability of malignancy

For the experiments where relevant, AUROC was used as a metric, as it is broadly used in the field due to its robustness against unbalanced evaluation sets.

The end2end models, namely the pure Resnet, pure VGG and the hybrid model, that were trained by (Shen et al., 2019), were tested on the CBIS-DDSM test set. For each *m* feature, the AUROCs were computed when the particular feature was excluded from the whole subset. This resulted in AUROCs on the whole test set and on *m* subsequent subsets for each model.

A simple ensemble method was made by averaging the posterior probabilities of each model and computing the AUROC of that. Moreover, each model was evaluated a second time with **input augmentation**, where each input image was flipped horizontally and vertically before evaluation.

The performance of radiologists is included, too. Here, the AUROC is computed using the ROI-level assessments with the following logic: The radiologist predictions for each whole image is malignant, if any ROI of the image has a BI-RADS 4, 5 or 6 assessment, and benign otherwise. Since the limit value of BI-RADS 4 corresponds to $3 - 94\%$ suspicion of malignancy (Liu et al., 2019), there is a high degree of uncertainty within the assessment. Therefore, the AUROC scores computed from the BI-RADS are not used to directly evaluate radiologists, but rather to assess if they are affected similarly to the end2end models, when hidden features cases are removed.

Prediction probability heatmaps were generated for examples of false-positive and false-negative cases with each hidden feature present. This was done by exploiting the intermediate patch classifier step of the end2end model architecture. A trained VGG19 patch classifier with a patch size of 224 and a stride of 32 was swept through each input image, outputting prediction probabilities for each patch of the image. The predicted probabilities correspond to the five classes, that the patch classifier was trained to predict: *background, benign-calcification, malignant-calcification, benign-mass* and *malignant-mass*. Each 2D array of prediction probabilities, excluding the background class, was turned into a heatmap by overlaying the input image with red or green depending on the predicted label (green for benign, red for malignant). The true ROI labels were also laid over the original image, as a means of comparison.

Lastly, **a multiple linear regression** was computed to search for correlations between the hidden features and the model posterior probability of malignancy. The analysis was conducted on the malignant cases and benign cases separately to avoid the effect of colinearity between the true label and the predicted probabilities.

6.2 Experimental setup: INbreast

As a way of evaluating data set characteristics as hidden features, the secondary INbreast data set was used. As noted previously, the INbreast set has some different characteristics compared to the CBIS-DDSM set. Mainly, it was made of full-field digital mammograms, while the CBIS-DDSM set was made of digitized film mammograms. Beside that, the set contained no instances with text present, which was heavily present in the other set, and the set further has a different distribution of BI-RADS assessments.

An experiment was set up to evaluate the performance of the end2end models on INbreast before and after transfer learning with the same test set in both cases. Hyperparameter-tuning was done at a small scale on the validation set. Adam optimizer was chosen with an initial learning rate of 0.0001, weight decays of respectively 0.001 and 0.01. Although 50 epochs were chosen, early stopping stopped the training at the 31st epoch.

6.3 Hardware Specification

Model evaluation was done on nodes of the ITU HPC cluster with 8 cores, 32GB memory and 1x RTX 2070 GPU, in an environment with Tensorflow 1.15.2, Keras 2.0.8 and Python 3.7.4.

6.4 Results

The results seen in **Table 4** show for radiologist assessments and for each of the end2end models, the AUROC score for the subsets of the CBIS-DDSM test set. The table shows a broad *decrease* in AUROC across the end2end models, when cases with text present (F1) are removed from the test set. Meanwhile, there is a broad *increase* in performance, when cases with a scar marker (F4) are removed.

The biggest performance gap occurs when text cases are removed before evaluating the ensemble model with augmented inputs. Here, the AUROC decreases by 10% points. On the other hand, the biggest single increase in performance occurs, when cases with nipple marker (F3) are excluded from the *Resnet* model. The two overall best performing models seem to be the hybrid and the ensemble model, both with augmented inputs.

Table 5 specifies the performance of the ensemble (the best performing model) further into BI-RADS-level. The table shows, that the ensemble model consistently performs high on BI-RADS 1 & 2, and conversely, it performs the worst on BI-RADS 0.

Figure 6 (bottom of document) shows heatmaps of prediction probabilities from sweeping a patch classifier on examples of false-positive cases from CBIS-DDSM test. For most of the cases, the patch classifier does not assign a high probability to any of the classes on patches that overlap the hidden feature. This is with the exception of the nipple marker in *2(a)*, where the benign calcification heatmap is highlighted on top of the nipple marker. Similar results are seen in **figure 7** for the false-negative cases.

Table 6 shows the results from multiple linear regression on all malignant cases and all benign cases. The linear models based on the hidden features explain respectively 5% and 18% of the total variation in the model predictions. In the benign cases, a significant, negative correlation of 0.22 was found for the text feature ($p < 0.05$) and a positive correlation of 0.33 for the "ruler" feature ($p < 0.05$). Lastly, in the malignant cases, a negative correlation of 0.3 was found for the scar marker feature ($p = 0.005 < 0.05$). The latter means that for malignant cases, the model assigns 30 (± 11) percentage points lower probability of cancer to cases with a scar marker.

Table 7 shows the performance of the three vanilla end2end models on a new data set, INbreast, before transfer learning, and the hybrid model after transfer learning. The former achieved slightly better than random results, while the latter achieved an AUROC of 78.9%.

7 Discussion

7.1 Interpretation of Results

Overall, there were signs that cases with particular hidden features confused the models, and signs of the opposite for other features, although not a proof of causality. For the **text feature**, there was a decrease in performance, when they were removed from the evaluation of the models on the CBIS-DDSM test set. This means, that the models generally had an easier time predicting instances with text present in the set. At the same time, a negative correlation was found between text cases and predicted probability of malignancy on benign cases. This indicates specifically, that the hybrid model found it easier to predict true-negatives on cases with text.

Exclusion of the **"ruler" feature** showed varying results in the evaluation. On the benign cases, the hybrid model predicted 33 (± 5.2) percentage points higher probability of malignant. In other words, the cases with the "ruler" feature likely had a higher false-positive rate.

Looking at the **nipple marker feature**, there was again a mix in how the models were affected, when the feature was removed, and there was found no significant linear correlation between the feature and model prediction.

Removing the **scar marker feature** resulted in an increase or stagnant performance for all the models. For malignant cases, the shown correlation indicates that the false-negative rate for cases with the scar marker was higher, because the model predicted the true malignant cases with a lower probability.

Although correlations were found, the results are inconclusive in finding what *causes* the changes in model predictions on cases with hidden features. It may be, that cases with certain hidden features simply have a different level of difficulty due to

Radiologist And Model Performance on CBIS-DDSM Test Set With Subsets Removed									
	Rad.	Resnet	VGG	Hybrid	Ensemble Avg.	Resnet*	VGG*	Hybrid*	Ensemble Avg.*
	AUROC								
None	66.1%	69.7%	75.5%	72%	76.8%	72.5%	78%	73.4%	78%
F1	55.9%	65.6%	67%	63.7%	68.7%	67.9%	68.2%	62.1%	68%
F2	66.9%	69.3%	75.9%	74.6%	77.6%	71.9%	78%	75.5%	78.4%
F3	65.8%	70.8%	73.9%	70.5%	75.6%	73%	76.3%	71.8%	76.5%
F4	66%	69.7%	76.2%	72.8%	77.4%	72.5%	78.8%	74.3%	78.6%

Table 4: Area Under ROC Curve for radiologists (*Rad.*) and end2end models on the official CBIS-DDSM test set, including metrics when subsets with each hidden feature are removed individually in the evaluation. **F1** are cases with text present, **F2** are "ruler" objects, **F3** are nipple markers and **F4** are scar markers. Asterix (*) indicates that the input is augmented before model evaluation.

Ensemble Average* Performance on CBIS-DDSM Test Set Per BI-RADS						
BI-RADS	All	0	1 & 2	3	4	5
	AUROC					
None	78%	47.7%	97.6%	63.3%	67.8%	56.5%
F1	68%	55.7%	n/a	61.3%	63.4%	66.7%
F2	78.4%	47.7%	98%	64.6%	68.4%	64.9%
F3	76.5%	47.7%	92.8%	55.1%	68.9%	56.3%
F4	78.6%	49.2%	97.6%	63.3%	68.2%	57.4%

Table 5: Area Under ROC Curve per BI-RADS for the ensemble of model averages on the official CBIS-DDSM test set, including metrics when subsets with each hidden feature are removed individually in the evaluation. **F1** are cases with text present, **F2** are "ruler" objects, **F3** are nipple markers and **F4** are scar markers. *Input images are augmented before evaluation.

other factors that affect the model predictions on these cases.

Looking at table 4, the estimated radiologist performance decreases by 10.2 percentage points, when text cases are removed, which is similar to all the end2end models. This indicates, that the cases with text present simply *are* easier to predict, even by humans, and hence we cannot conclude that the hidden feature is the cause of the decline in model performance. On the other hand, the cases with scar marker affected the performance of the models and the radiologists reversely. While most models increased in performance without the scar marker cases, the radiologist performance decreased by .1 percentage points. Although not conclusive, this could indicate, that the scar marker cases have some underlying difficulty for the models to predict, which does not affect humans.

The heatmaps in figure 6 & 7 indicate that the patch classifier predicts "background" (irrelevant) on most patches that overlap a hidden feature. This means that the first step of the 2-step architecture in the end2end models assesses the hidden features as non-indicative of neither benign nor malignant. It could potentially be inferred from this, that the end2end models as a whole are not affected by the features, however, such a conclusion is hard to draw, as there may be a lot more going on under the hood of the top layers, before the final whole image prediction output. For starters, recall how the top layers in the architecture were put on top of the final layers of the patch classifier - rather than the outputs. This was done by the authors (Shen et al., 2019) to avoid a bottleneck architecture, where information from the patch classifier is reduced to few dimensions in the outputs. For this reason, the whole image classifiers may still be triggered by the hidden features in ways unknown, even if the patch classifier in itself does not trigger.

The INbreast experiment shows a clear difference in performance before and after transfer learning on a new mammography data set. This is by no means new knowledge - that transfer learning can increase the performance of models on unseen

Multiple Linear Regression On Hidden Features And Posterior Probabilities				
Hidden feature	Coefficient	Std. error	t	$P > \|t\|$
Malignant cases: $R^2 = 0.049$				
F1	-0.0308	0.043	-0.711	0.478
F2	0.0762	0.050	1.529	0.127
F3	0.0745	0.049	1.528	0.128
F4	-0.3020	0.106	-2.849	0.005
Benign cases: $R^2 = 0.177$				
F1	-0.2215	0.040	-5.480	0.000
F2	0.3297	0.052	6.295	0.000
F3	-0.1015	0.061	-1.652	0.099
F4	0.0390	0.115	0.339	0.735

Table 6: Multiple linear regression results on CBIS-DDSM test set for malignant cases and benign cases separately. Hidden features are the predictors and the posterior probability *p(malignant)* of the hybrid model is the response variable. **F1** are cases with text present, **F2** are "ruler" objects, **F3** are nipple markers and **F4** are scar markers.

Model Performance on INbreast Test			
Before transfer			*After transfer*
Resnet	VGG	Hybrid	Hybrid
58.2%	54.4%	51.9%	78.9%

Table 7: AUROC performance of end2end models trained on CBIS-DDSM train set and tested on INbreast test set. *Hybrid after transfer* is the hybrid VGG and Resnet model that has been transfer learned on a train set of INbreast before testing.

data sets. However, it does set the stage for other, new experiments mentioned as future work in the following section.

7.2 Future Work

Given the low sample size of $n = 645$ in the CBIS-DDSM test set, a more comprehensive study would be necessary to fully conclude the effects of the selected hidden features on machine learning models in mammography. A more robust approach would be to utilize the full data set of 3.103 whole mammograms, re-train and fine-tune several models on different train-validation-test splits, and average their performances on the test sets. Such an experiment would reduce the likelihood that the hidden features in a certain test set are simply tougher or easier by chance.

Work is still needed to evaluate the effects of the hidden features not found in the CBIS-DDSM test set, such as pacemakers, implants and demographics. There was one case with a pacemaker present in the whole CBIS-DDSM set (including train), meaning a study of such a feature would require a new data set. The Arabic KAU-BCMD data set (Alsolami et al., 2021) was considered in this project, as it contained several cases with implants (example in figure 8), but it was disregarded due to time constraint.

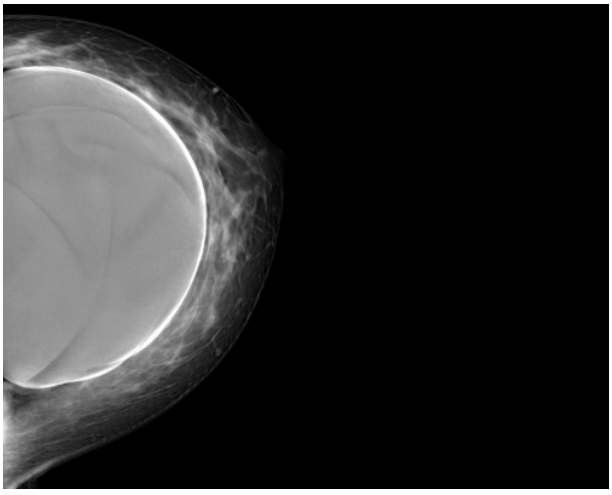


Figure 8: Example mammogram of a patient with implants from the KAU-BCMD data set.

Following the positive results of the transfer learning experiment, it would be valuable to transfer learn a neural network on a train set of KAU-BCMD and evaluate if such a model differs in per-

formance on cases with implants.

Methods of explainable AI such as Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2016) could be useful tools to give a visual explanation (through saliency maps) of the decisions made by the neural network, in a similar fashion to the patch classifier heatmaps in this report.

8 Conclusion

This study explored the effects of hidden features on machine learning models, particularly Resnet and VGG CNNs, in breast cancer detection. Models built on an end-to-end approach, with a whole image classifier on top of a patch classifier, were tested on the official CBIS-DDSM test set. The data set was annotated with several hidden feature types, resulting in 471 cases with text, 97 with a "ruler" object, 86 with nipple marker and 18 with scar marker. Cases with text present were easier for the models to predict, but likewise for estimated radiologist performance, meaning it could be due to those cases inherently being easier to predict in the data. Scar cases were harder for the models to predict, and seemingly had a higher false-negative rate. Visual analysis of the intermediate patch classifier did not indicate an effect of the hidden features on the decision-making. More work on bigger sample sizes is needed to be conclusive. Suggested further work includes:

- Training and testing models on different data splits, and evaluating the average performance, as a means to show variability in the results. Preferably on a bigger data set.
- Interpreting the decision-making of models through explainable AI, such as Grad-CAM.
- Transfer learning on other data sets with new hidden feature types, like pacemakers and implants, and evaluating model performance on such cases

If more work is done in the field of hidden features and machine learning failures in breast cancer detection, it could increase the robustness of models before clinical usage. Such an effect would mean better health to more people.

9 Acknowledgement

I would like to thank Veronika Cheplygina for being an inspirational supervisor, and for being the gateway for me to work on a project in an exciting field. I would further like to thank Lottie Rosamund Greenwood for providing outstanding support on the ITU HPC cluster.

References

Asmaa S Alsolami, Wafaa Shalash, Wafaa Alsaggaf, Sawsan Ashoor, Haneen Refaat, and Mohammed Elmogy. 2021. King abdulaziz university breast cancer mammogram dataset (kau-bcmd). *Data*, 6(11):111.

Rama S Ayyala, MaryAnn Chorlton, Richard H Behrman, Phyllis J Kornguth, and Priscilla J Slanetz. 2008. Digital mammographic artifacts on full-field systems: what are they and how do i fix them? *Radiographics*, 28(7):1999–2008.

Sai Batchu, Fan Liu, Ahmad Amireh, Joseph Waller, and Muhammad Umair. 2021. A review of applications of machine learning in mammography and future challenges. *Oncology*, 99(8):483–490.

Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. 2018. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424.

Tomas Budrys, Vincentas Veikutis, Saulius Lukosevicius, Rymante Gleizniene, Egle Monastyreckiene, and Ilona Kulakiene. 2018. Artifacts in magnetic resonance imaging: how it can really affect diagnostic image quality and confuse clinical diagnosis? *Journal of Vibroengineering*, 20(2):1202–1213.

Cancer.org. 2022. Cancer survival rates.

L Chaloeykitti, M Muttarak, and KH Ng. 2006. Artifacts in mammography: ways to identify and overcome them. *Singapore medical journal*, 47(7):634.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. 2021. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. arxiv 2015. *arXiv preprint arXiv:1512.03385*.

Thomas Henn, Yasukazu Sakamoto, Clément Jacquet, Shunsuke Yoshizawa, Masamichi Andou, Stephen Tchen, Ryosuke Saga, Hiroyuki Ishihara, Katsuhiko Shimizu, Yingzhen Li, et al. 2021. A principled approach to failure analysis and model repairment: Demonstration in medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 509–518. Springer.

Kræftens Bekæmpelse. 2020. Statistik om brystkræft.

Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. 2017. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific data*, 4(1):1–9.

Gang Liu, Meng-Ke Zhang, Yan He, Yuan Liu, Xi-Ru Li, and Zhi-Li Wang. 2019. Bi-rads 4 breast lesions: could multi-mode ultrasound be helpful for their diagnosis? *Gland surgery*, 8(3):258.

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

Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso. 2012. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248.

RR Selvaraju, M Cogswell, A Das, R Vedantam, D Parikh, and D Batra. 2016. Grad-cam: visual explanations from deep networks via gradient-based localization. 2016. *arXiv preprint arXiv:1610.02391*.

Li Shen, Laurie R Margolies, Joseph H Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. 2019. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9(1):1–12.

Yoel Shoshan, Ran Bakalo, Flora Gilboa-Solomon, Vadim Ratner, Ella Barkan, Michal Ozery-Flato, Mika Amit, Daniel Khapun, Emily B Ambinder, Eniola T Oluyemi, et al. 2022. Artificial intelligence for reducing workload in breast cancer screening with digital breast tomosynthesis. *Radiology*, page 211105.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Xiaoqin Wang, Gongbo Liang, Yu Zhang, Hunter Blanton, Zachary Bessinger, and Nathan Jacobs. 2020. Inconsistent performance of deep learning models on mammogram classification. *Journal of the American College of Radiology*, 17(6):796–803.

Kevin Wu, Eric Wu, Yaping Wu, Hongna Tan, Greg Sorensen, Meiyun Wang, and Bill Lotter. 2019a. Validation of a deep learning mammography model in a population with low screening rates. *arXiv preprint arXiv:1911.00364*.

Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzebski, Thibault Févry, Joe Katsnelson, Eric Kim, et al. 2019b. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE transactions on medical imaging*, 39(4):1184–1194.

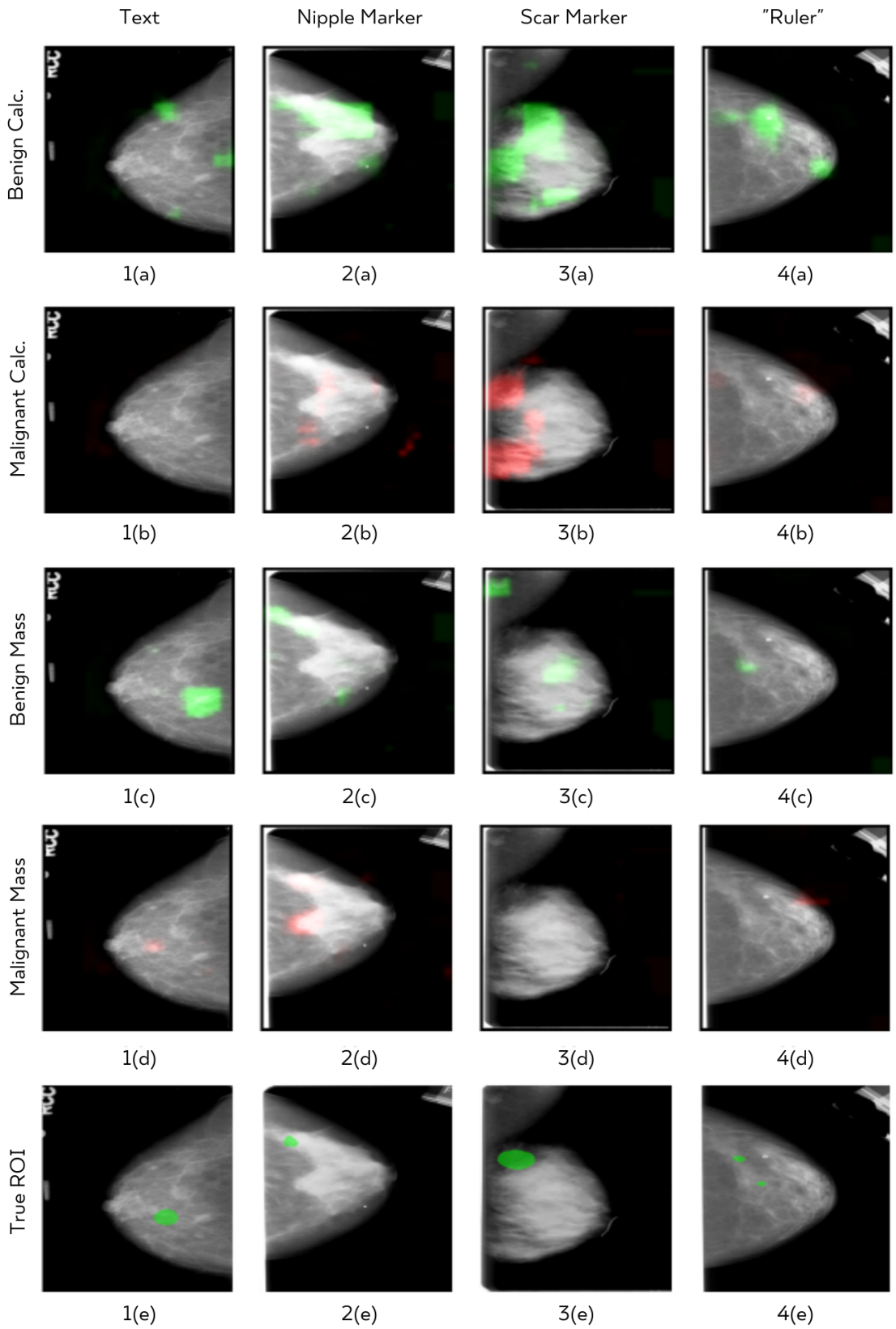


Figure 6: Examples of heatmaps from sweeping a trained VGG patch classifier over cases of **false positive** images in the CBIS-DDSM test set. One case for each hidden feature.

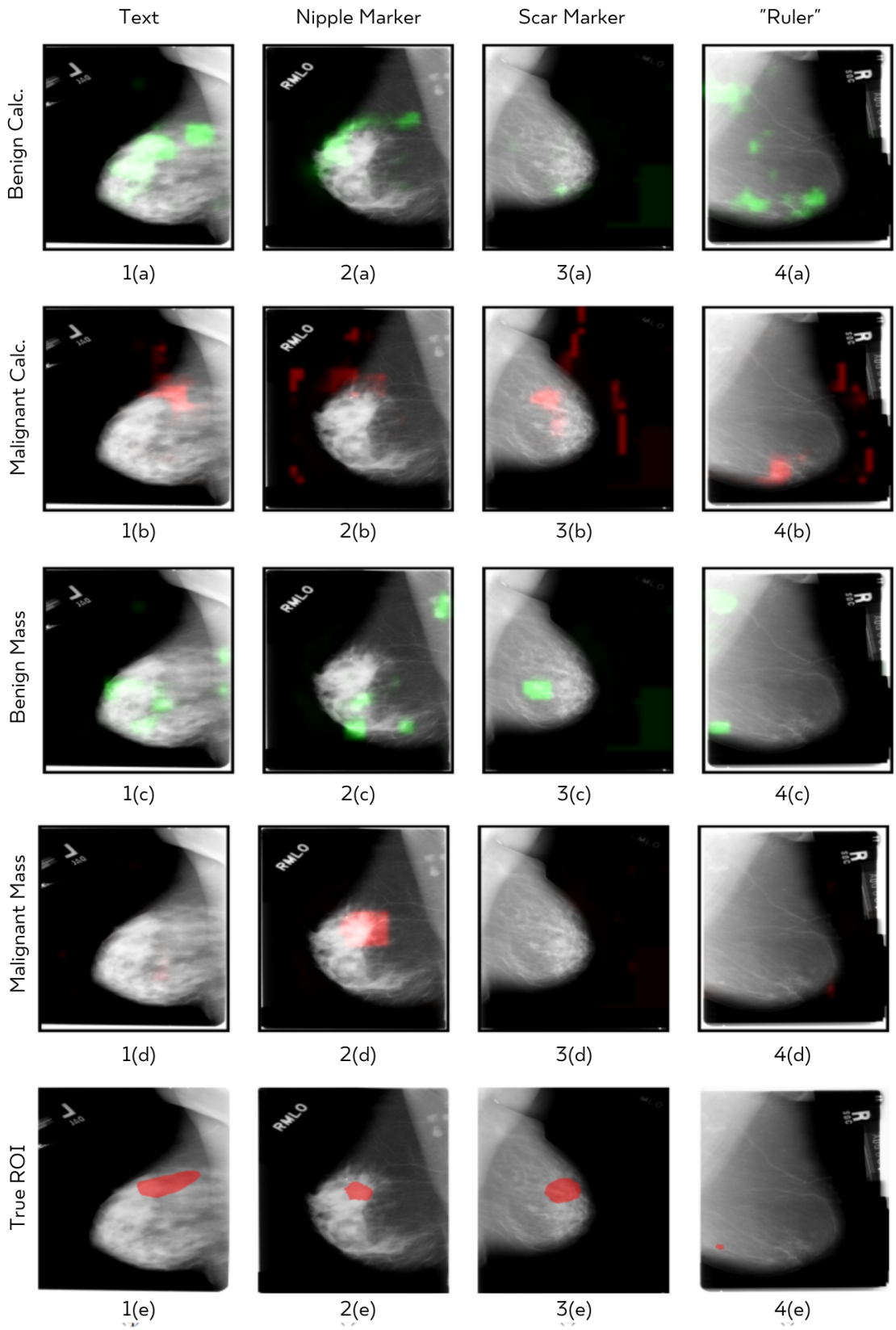


Figure 7: Examples of heatmaps from sweeping a trained VGG patch classifier over cases of **false negative** images in the CBIS-DDSM test set. One case for each hidden feature.