Appendix

A Image selection procedure flowchart

A flowchart detailing the data selection process for the creation of CSAW-M.

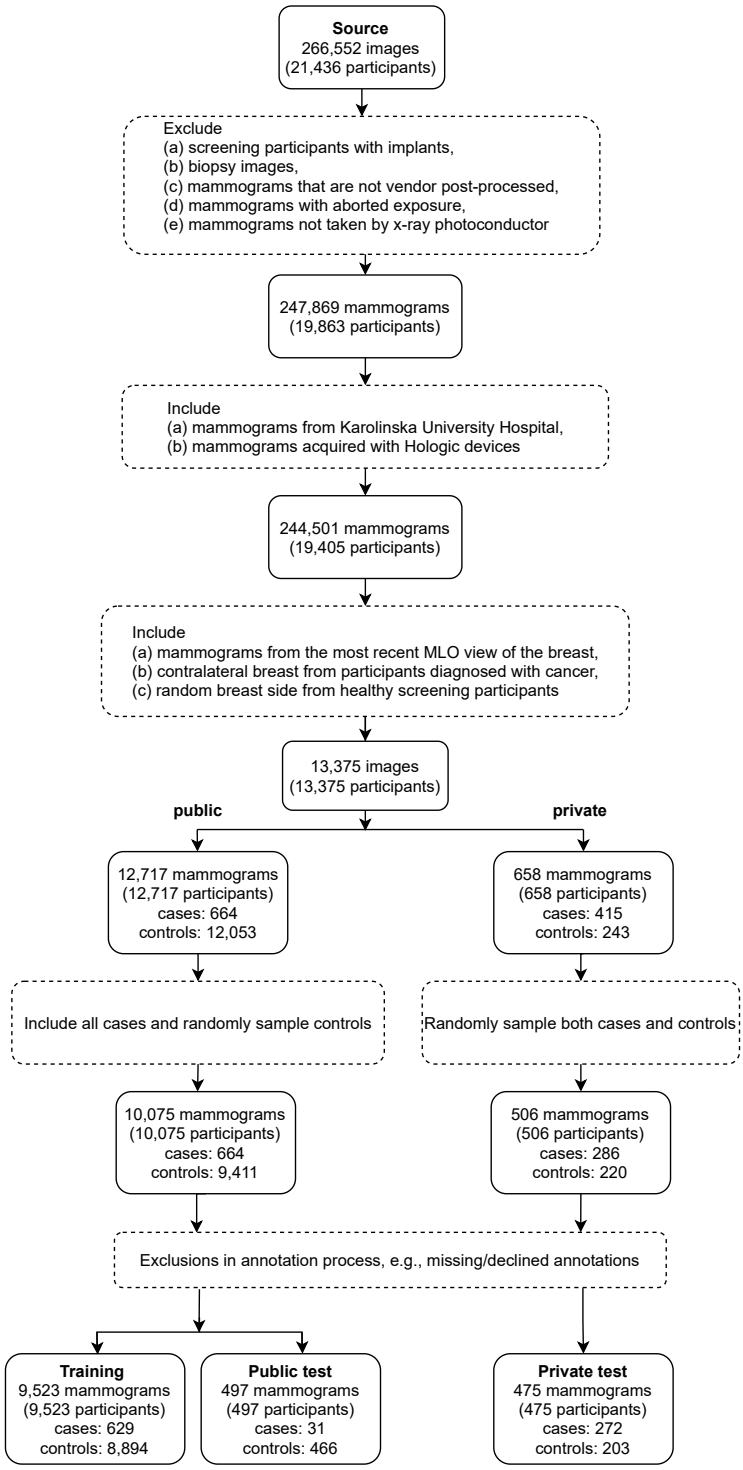
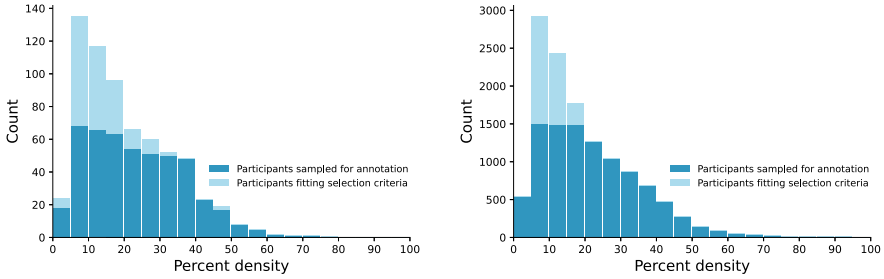


Figure 7: Data selection flowchart.

B Sampling screening participants and pre-processing of images in CSAW-M

Sampling screening participants for CSAW-M by percent density. In order to have images with diverse breast densities, we uniformly sampled images with intermediate densities while keeping all images in the tails. The percent densities are calculated with the publicly available software Libra [10]. Figure 8 shows the resulting distribution of breast density in the public and private portions of CSAW-M.



(a) The private test set

(b) The public training and test sets

Figure 8: Distribution of breast percent density for different splits of CSAW-M.

Finding and removing the area in the image that contains text. We remove the text in the image using the OpenCV library [32]. We try to find the contour which includes text in the image, and then replace the pixels within that contour. Using a binarized copy of the image (in which each pixel greater than 0 is mapped to 255 and pixels with value 0 remain 0, resulting in a totally black and white copy of image), we first extract all available contours. Note that the letters themselves are among the contours. We ignore the largest contour in the image - which corresponds to the breast and consider other available contours as candidates. This ensures that the breast remains intact. Since we know that the letters in the image are spatially close to each other, we apply a dilation transformation to the image so we could combine the little adjacent contours in the image. This transformation essentially combines smaller contours corresponding to letters into a larger contour. We also ignore all the contours below the breast, since we know beforehand that the text will never be there. Among the remaining contours, we finally only consider the ones whose area is at least greater than a threshold (which we found by trial and error). If there are multiple contours remaining, we choose the one that is closest to the top-right corner of the image, since we already know the text is almost always around there. This will result in the contour containing the text in the image. Once the contour including the text is found, we use its coordinates and replace the values inside it in the original image with the minimum pixel value that is available in the original image (which is usually black, corresponding to air).

C Details on the annotation procedure

Annotation tool. We developed an annotation tool and distributed it to the experts for annotating CSAW-M. The tool displayed two mammogram for a *pairwise comparison* and asked the expert: *which image is harder to be certain there is no tumor?*. The interface is shown in Figure 9. In order to make the comparison easier for the annotators, images are flipped if necessary so that both breasts point to the same direction, making the assessment easier. We always showed the *query image* on the left and the *reference image* on the right. We asked the annotators to press "1" is they evaluate that the left image is harder to assess (exhibiting a higher masking level), "2" if the right image is harder to assess, and '9' if they see no discernible difference between them. At any point in time, annotators were provided the option to *discard* the current pairwise in case they could not make a reasonable judgement, *e.g.* because the query image is distorted, it does not contain the whole breast etc. The tool was used for both binary and ternary search without any changes to the user experience.

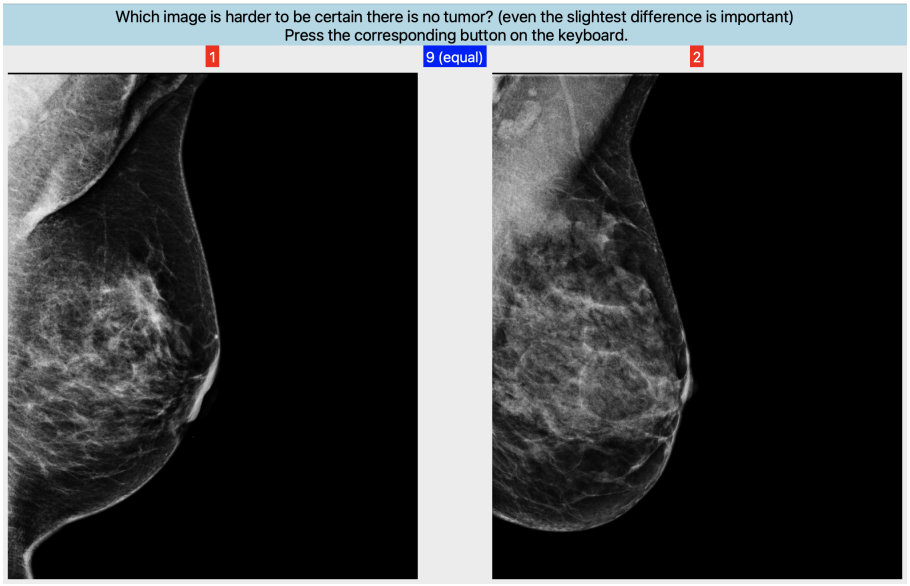


Figure 9: Interface of our annotation tool. See text for details.

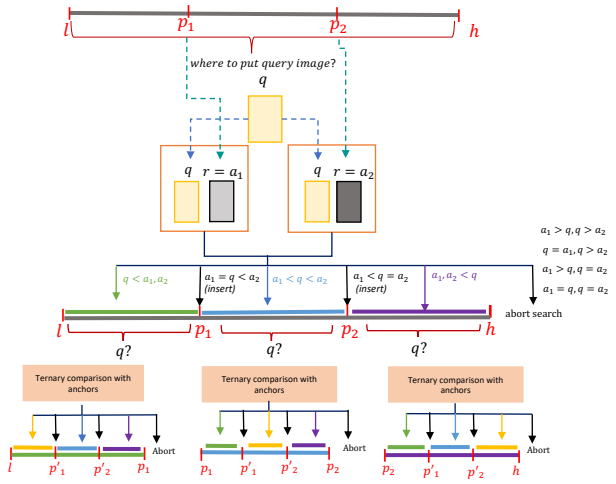


Figure 10: Illustration of the pairwise comparisons used in our ternary search, where the query image is compared against two anchors from the sorted list. Based on the assessment from the annotator, either the ternary search interval gets updated, insertion happens, or the search is aborted.

Technical details about the ternary comparisons for the private test set There are 9 possibilities for how the query image q could be assessed against anchors a_1 and a_2. Figure 10 illustrates different possibilities and show the outcome of each of them. We consider the two consecutive comparisons *consistent* if: 1) $q > a_1, a_2$ 2) $q < a_1, a_2$ or 3) $a_1 < q < a_2$. In the case of consistent comparisons, the search interval will be updated accordingly, and the search will continue to its next level. We directly insert q at position p_1 if $a_1 = q < a_2$ and we insert at position p_2 if $a_1 < q = a_2$. All other possibilities are considered to be *inconsistent*, in which case the search is *aborted*, and we start the search from the beginning for a new query image. At the end of the annotation procedure, we asked the experts to make new attempts to insert the aborted images into the list. Since the sorted list would continuously get updated after inserting new images, the new attempts would usually result in successfully inserting the images whose search was previously aborted. Note that in the new attempts, the search process for inserting the query images would begin from scratch.

Technical details about the binary search for the private test set Given the search interval $[l, h]$, the image lying at position $p = (l + h)/2$ would be selected as the *reference* image. The search interval is updated based on the result of the pairwise comparison. If $q > r$ the search interval would change to $[p + 1, h]$, if $q < r$ the search interval would change to $[l, p - 1]$, and if $q = r$ the image would get directly inserted at position p.

D Technical details on the models

Implementation details We used the PyTorch framework [33] for training our networks. We also used the scikit-learn [34] and SciPy [35] libraries for computing the metrics we used in our experiments. Training each model was done on a single NVIDIA Quadro RTX 8000, taking ∼2 hours to complete.

Calculating the continuous masking score We can aggregate the output probabilities of a model into a single continuous score that we call *masking score*. For the *one-hot* model, we simply compute the score as a weighted average of probabilities as:

$$s = \frac{1}{8}(\sum_{l=1}^{8} p_l \cdot l) \tag{1}$$

where p_l represents the probability that the input image belongs to masking level l.

To get continuous scores from the *multi-hot* model, we first need to convert the cumulative probabilities that each output head produces to actual masking level probabilities p_l. Recall that the multi-hot model trained for ordinal classification with 8 classes would have 7 output heads. As mentioned in the main text, each output head models an individual binary classifier. For an input image x with true label L, each output head *independently* produces $P_k = p\,(L > k)$ where $k \in \{1, 2, .., 7\}$, denoting the (cumulative) probability that L exceeds k. Since in general each output head is an independent binary classifier, the output probabilities are not guaranteed to be *monotonic* (although most often they are), so we apply a small trick to make them monotonic, *i.e.,* for each $k' > k$, it holds that $P_{k'} <= P_k$. This is implemented as in the code snippet below.

```
def make_monotonic(cdf_list):
    monotonic = []
    for i in range(7):
        max_cdf = max(cdf_list[i:])
        monotonic.append(max_cdf)
    return monotonic
```

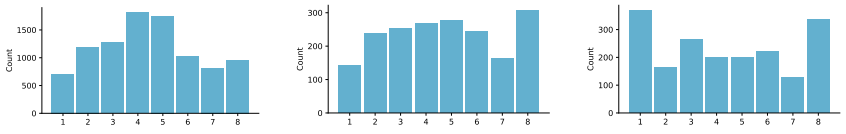
Once the probabilities are made monotonic, individual masking level probabilities could be calculated by subtracting consecutive cumulative probabilities such that for each $k' = k + 1$ it holds that $p_k = P_k - P_{k'}$, as shown in the following code snippet:

```
def make_probs(lst):
    extended = deepcopy(lst)
    extended.insert(0, 1)  # cdf: 1 in the beginning
    extended.append(0)  # cdf: 0 at last
    probs = []
    for i in range(0, len(extended) - 1):
        probs.append(extended[i] - extended[i + 1])
    return probs
```

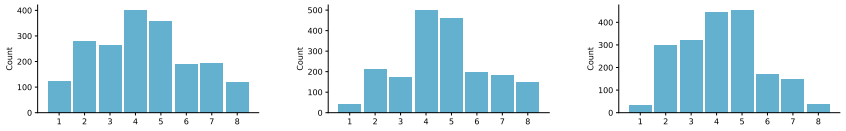
Once we have evaluated individual masking probabilities, we could use the same formula as in Equation 1 to calculate the continuous score.

E Distribution of ground-truth and annotations from each expert

In Figure 11 and Figure 12 we provide the distributions of masking level annotations for each expert on the public training and test sets. The distribution of ground truth annotations is also provided. The ground truth for each test image is the median of expert annotations.

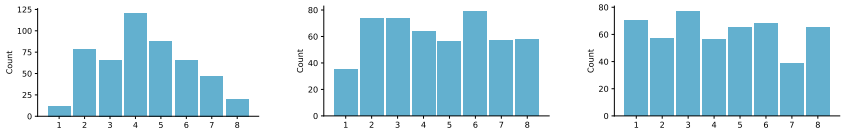


(a) GT (b) Expert 1 (c) Expert 2

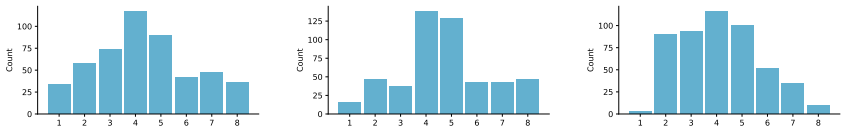


(d) Expert 3 (e) Expert 4 (f) Expert 5

Figure 11: Distribution of ground-truth and annotations from each expert on public training set.



(a) GT - Median (b) Expert 1 (c) Expert 2

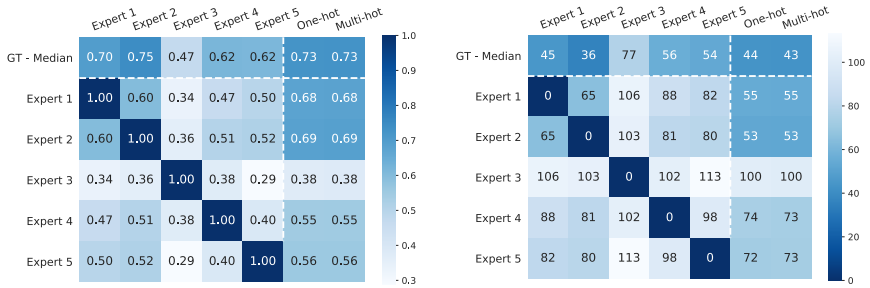


(d) Expert 3 (e) Expert 4 (f) Expert 5

Figure 12: Distribution of ground-truth and annotations from each expert on public test set.

F Ordinal classification of masking potential on private test set

Model and expert agreement on the private test set is shown in Figure 13 with Kendall's τ_b and *AMAE*. Similar conclusions can be drawn in private test set as that of public test set in Figure 5.



(a) Kendall's τ_b (higher is better) (b) AMAE (lower is better)

Figure 13: Expert and model agreement on private test set. We perform five runs on our one-hot and multi-hot models and report the mean.

G Identification of large invasive cancer

Figure 14 shows the odds ratio on large invasive cancer with public and private test sets combined. The results are less promising compared to the odds ratio on interval cancer and CEP, shown in Figure 6.

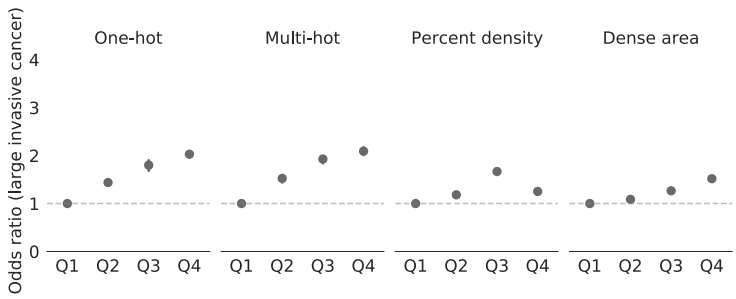


Figure 14: Odds ratio on large invasive cancer with public and private test sets combined.

H Variations in masking levels

In Figure 15, violin plots show the distribution of percent density as a proxy of variation within each masking level, grouped by expert. We can see that experts 1 and 2, whose masking-level distribution was almost uniform (as seen in Figure 12), show high variability for high masking levels. This can be explained by the heavy tail of the percent density distribution even after undersampling (as described in Figure 8). Uniform binning would result in more images with diverse percent density (from the tail of the distribution) to be included in their high masking levels. Moreover, the median of the violin plots changes more significantly for these experts as in higher masking levels.

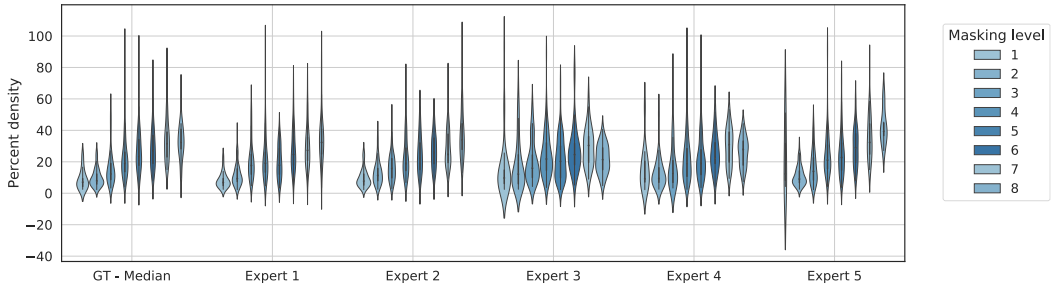


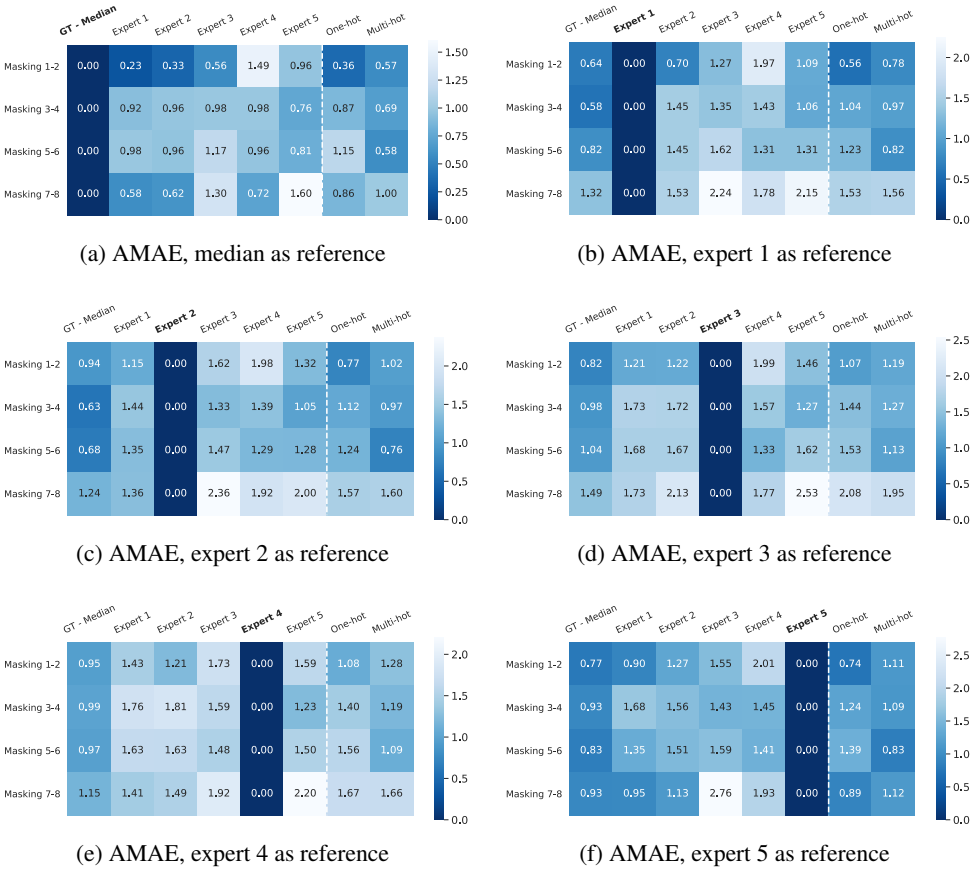
Figure 15: Variations in different masking levels using percent density as a proxy on public test set.

I Agreements in masking levels

In Figure 16, we provide annotator agreements across different masking levels. Here, we divided masking levels into 4 non-overlapping groups: (1, 2), (3, 4), (5, 6), and (7, 8). For each of these groups, we individually considered each expert (and GT-median) as the reference, and evaluated how other experts/models agree with respect to the selected reference in terms of the *AMAE* measure. For example, in the top row of Figure 16a we consider the GT-median as reference, and compute the *AMAE* of the other experts. A low value indicates they agree with the GT-median in discriminating masking level 1 from masking level 2. We repeated this process for the other groups to fill in the remaining rows.

From Figure 16, we can observe that in general disagreement is high in higher masking levels. This trend is especially obvious for experts 3, 4 and 5. The experts tended to agree best when discriminating between middle masking levels (3,4) and (5,6). Performance for the low masking levels (1,2) was interesting, as certain experts tended to agree with each other well (experts 1 and 2)

while others disagreed strongly. Finally, the trained models also exhibit this trend, and it can be seen that generally the models and the median agree more than other experts.



(a) AMAE, median as reference

(b) AMAE, expert 1 as reference

(c) AMAE, expert 2 as reference

(d) AMAE, expert 3 as reference

(e) AMAE, expert 4 as reference

(f) AMAE, expert 5 as reference

Figure 16: Expert and model agreement in different masking levels on public test set. We perform five runs on our one-hot and multi-hot models and report the mean.

J Dataset accessibility

Our dataset can be accessed using the DOI: 10.17044/scilifelab.14687271. Access to the dataset files are given upon agreeing to the terms and sending a request. We note that the dataset webpage is self-contained, that is, all the data and metadata files needed for the user to understand how to use the data are available in the same place.

K Hosting, licensing, and maintenance plan

The dataset is hosted by https://scilifelab.figshare.com/. This interface has restricted access where users must submit their request, after which the access to the actual files will be granted. We have been in contact with staff from SciLifeLab Data Repository and have received a letter from them (available on the final page), supporting the hosting of our data. We will actively check for requests made to the data. In case any change is required to be made to the dataset, we will use the *versioning* functionality provided by the repository. Through this functionality, all the previous versions of the data will still be available. This change will be communicated clearly to the users and will also be reflected on the dataset landing page.

The site contains instructions on how to request the data. The data is self-contained – enabling users to easily understand the content and organization of the files using the provided metadata file. Please visit the dataset webpage for license and terms[9].

Planned service. SciLifeLab Data Repository will provide an infrastructure to run AI models on the repository using Kubernetes and Docker. Although the plan is not definite yet, we will try to use the provided infrastructure so users could evaluate models trained to estimate masking level using the CSAW-M private test set. This will allow researchers to evaluate their models in a less biased setting, as the (planned) evaluation server will limit the number of times users can submit their models, so models cannot overfit the private test set.

L Author statement of responsibility

The authors are responsible for violations of privacy and data ownership laws pertaining to the distribution of CSAW-M, including:

1. violations of data protection law, specifically GDPR;
2. violations of the terms of approvals from the Ethical Review Board (EPM 2021-01030).

The authors bear responsibility for taking the necessary technical and organizational measures to protect the data from re-identification.

M Documentation framework and intended use of the dataset

Here we provide documentation for the CASAW-M dataset. We use the Datasheets for Datasets framework [16] to document our dataset.

MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
The main goal in creating this dataset was to enable the development of models capable of identifying screening participants in mammography screening whose mammograms cannot easily be assessed due to a high level of mammographic masking, a phenomenon that occurs when potential cancer could largely be obscured by the surrounding tissue in the breast. As a result, breast cancer in these participants is more likely to be missed during regular mammography. More sensitive imaging technologies such as MRI are too costly to be provided for all participants visiting a clinic. Due to the large number of mammography images that are taken at clinics, there exists a need to develop an AI model that could help identifying screening participants in higher needs of MRI. To develop such an AI model, we noticed a lack of mammographic images containing assessment of masking level directly made by expert radiologists. Although public mammographic datasets exist, none of them exactly contains direct potential of masking in mammograms assessed by radiologists. Our aim was to fill the gap by collecting a dataset that merely focuses on mammographic masking. CSAW-M helps to automate identifying of low- and high-masking mammograms.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The dataset was created in a joint collaboration of researchers from KTH Royal Institute of Technology, Karolinska Institutet, Karolinska University Hospital, and S:t Görans Hospital in Stockholm.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.
This work was partially supported by MedTechLabs https://www.medtechlabs.se/, the Swedish Research Council (VR) 2017-04609, and Region Stockholm HMT 20200958.

[9]The dataset can be found with this DOI: 10.17044/scilifelab.14687271

Any other comments?
None.

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
The dataset comprises mammographic images, together with metadata that is provided as CSV files. The metadata includes masking potential labels collected from five experts, image acquisition parameters, clinical endpoints *i.e*. cancer attributes and density measures. More details can be found in the main paper.

How many instances are there in total (of each type, if appropriate)?
There are 10,020 screening participants in total, and each participant has 1 mammogram from the MLO view of the breast. 9,523 of the images are in CSAW-M training set with one annotation per image, while the rest 497 images are from a public test set where each image has annotations from 5 experts.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
CSAW-M is a subset of CSAW, a large population-level cohort of screening mammograms [25]. We exclude images that are: *a)* from patients with implants; *b)* biopsy images; *c)* mammograms that are not vendor post-processed; *d)* mammograms with aborted exposure; *e)* mammograms not taken by X-ray photoconductor; *f)* earlier mammograms when there are duplicates in the same exam. We sample screening participants with complete mammography exams taken in Karolinska University Hospital and with Hologic manufacturer. We sampled from the participants according to the procedure mentioned in Section 2 in a way that more mammograms with extreme density values are included (which are more clinically interesting), so the sampled mammograms are not necessarily representative of the larger set. This was done because mammograms in the tails of the percent density distribution (very dense or very fatty) are of the highest clinical interest.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Images in PNG format, certain DICOM acquisition attributes that can be used for preprocessing, clinical endpoints, and masking annotations from 5 experts (as detailed in Table 2). Training images have one annotation while test images have 5 annotations per image.

Is there a label or target associated with each instance? If so, please provide a description.
Yes, the labels are masking levels (from 1-8) of each instance annotated by 5 experts, together with certain clinical endpoints, *i.e*. interval or large invasive cancer.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
No, there is no missing information. The information is complete for all individual instances.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
Yes, the annotations are explicitly applied to the images, which were shown directly to the experts.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
Yes, we have recommended training and testing splits. We have benchmarked mammographic masking of cancer on suggested testing splits where there are 5 annotations per image (the median was chosen as ground truth), as opposed to the training set that contains 1 annotation per image. There is no recommended development/validation split. However, we have provided the cross-validation folds that we used when developing the models.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
Yes. Experts may have made wrong button clicks or errors in their comparisons. Similarly, there may be clerical errors matching patients with their clinical endpoints.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
The dataset is hosted by the `https://scilifelab.figshare.com/`. It has restricted access where users must submit their request, after which the access to the actual files could be granted. See Appendix K for more details.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
Yes, the dataset contains information related to the health status of individuals. The information has been reduced in order not to allow the identification of any individual. In our assessment, the dataset contains only de-identified information.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
No. Our dataset mainly contains mammography images, and there is nothing offensive, insulting, or threatening.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
Yes, the source dataset is extracted from a large cohort containing millions of mammograms, collected every 18 to 24 months from screening participants aged 40 to 74 in Stockholm county area. The dataset contains around 10,020 mammograms taken from 10,020 participants.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
In Sweden, individuals with female personal identity numbers are invited for mammographic screening. Our dataset contains mammogramgraphic screenings from screening participants 40 to 74 years of age. Racial information is not collected in Sweden.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
No, we have taken appropriate measures to ensure it is not possible. The measures include: (1) we removed all individual identifiers from the data, (2) we down-sampled the mammograms, (3) we removed all unnecessary acquisition attributes –DICOM headers–, (4) we simplified the continuous tumor size attribute to a binary outcome, and (5) we anticipated a gated release mechanism to approve users based on their information and project goals before granting access to the data.

Users are also required to explicitly agree not to attempt to de-identify any individuals from the dataset.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.
The cancer images in our dataset are accompanied with the clinical outcome of the screening corresponding to that image, *i.e*. whether they are diagnosed with interval or large invasive cancer. These attributes, however, are available in our dataset in a binary form and the screening participants are de-identified.

Any other comments?
None.

COLLECTION PROCESS

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
The source images of our dataset are in DICOM format with DICOM metadata. Each patient is linked to the Regional Cancer Registry to define clinical endpoints such as whether a screening participant was healthy or had been diagnosed with breast cancer. Mammograms were shown to five experts to assign masking annotations.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?
We designed a user interface through which we showed two images alongside each other and asked experts to do pair-wise comparisons and select the image that is harder to assess. It was validated to be bug-free by running several tests with experts before the collection process began.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
The data is sampled from CSAW as explained in Section 2. The sampling was done in way to include more mammograms with very low or very high percent density measure as these are the most clinically interesting images.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?
Researchers from KTH Royal Institute of Technology, Karolinska Institutet, Karolinska University Hospital, and S:t Görans Hospital in Stockholm were involved in the data collection. All participants were compensated for their time in the course of their normal research activities.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.
The mammograms were collected during regular mammography screening between 2008 and 2015 at Karolinska University Hospital. The creation of the CSAW-M dataset, including developing the annotation tool, receiving annotations from experts, cleaning data etc. was initiated in June 2020 and lasted until November 2020.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
The Regional Ethical Review Board in Stockholm has approved the research. Also, a dedicated agreement between Karolinska Institutet and KTH Royal Institute of Technology has been made to publish the data.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
The source mammograms were collected by Karolinska University Hospital, the clinical labels were collected by the Regional Cancer Center, and masking labels were collected by showing mammograms to five experts for annotation.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
No. The need for informed consent was waived by the Ethical Review Board.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
No. The need for informed consent was waived by the Ethical Review Board.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)
Not applicable.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
We consulted an expert in GDPR and dealing with personal data from Karolinska Institutet, and our concerns regarding privacy were cleared.

Any other comments?
None.

PREPROCESSING / CLEANING / LABELING

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
Data preprocessing was done in our baseline implementations. The source images of our dataset were DICOM files whose pixel values we saved as raw PNG images. Using DICOM metadata, we did preprocessing to generate PNG images. Please refer to Section 2 for more details about image preprocessing.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
The "raw" data was saved as PNG, and we also provide the preprocessing script that was used in our baseline implementation for reproducibility.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
We used Python standard libraries and the preprocessing script is available in the Github repo of the project: https://github.com/yueliukth/CSAW-M/.

Any other comments?
None.

USES

Has the dataset been used for any tasks already? If so, please provide a description.
Yes. The masking model, together with two other models that we developed to perform breast cancer risk prediction and cancer detection, are combined into a single comprehensive model. This clinical workflow is currently implemented at Karolinska University Hospital in a clinical study to help identify screening participants who are most likely to benefit from additional MRI screening.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
No.

What (other) tasks could the dataset be used for?
First, our dataset has annotations that are ordinally related and can be used to study ordinal classification or point-wise ranking tasks. Specifically, our public test set contains 5 annotaitons per image, which make it a useful resource to study human noise and bias. Moreover, our dataset which contains more than 10,000 mammograms, is significantly larger than other public mammography datasets (see Table 1 in the main paper). It can be used for pretraining deep learning models that would be used in other downstream tasks in a similar domain to mammography images (for more effective transfer learning). And last but certainly not least, we included clinical endpoints as our metadata, making it valuable in clinical studies. We have shown in the paper that our ResNet-34 models trained on estimating masking potential perform better than the breast density counterparts in identifying screening participants diagnosed with interval and large invasive cancers, without being explicitly trained for these tasks. This shows a great promise for the usefulness of our collected labels and motivates developing better models for estimating masking level.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
We are aware of the fact that biases exist inherently in our data collection, for the following reasons: *(a)* the data was extracted from a certain population, period and region, with certain manufacturers, *(b)* the annotations were made by radiologists from a certain region, *(c)* we randomly sampled screening participants and intentionally selected breasts that are denser or fattier which resulted in a distribution that is not representative of the real population. We note that clinical studies are crucially required before deploying models in any clinical processes.

Are there tasks for which the dataset should not be used? If so, please provide a description.
In the main article, we have noted that our dataset is not aimed for developing/evaluating cancer detection models, as the cancer images in CSAW-M are chosen to be *contralateral* to cancer laterality, *i.e*. the breast that does *not* contain tumor was selected (please refer to Section 2 for motivation).

Any other comments?
None.

DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
SciLifeLab Data Repository, who hosts our dataset, is currently relying on the Figshare service, but plans to move data to its own storage servers soon. This does not change availability of the dataset in any way, nor does it impose additional restrictions by any third party. SciLifeLab Data Repository is affiliated with KTH Royal Institute of Technology and the hosting of our dataset is guaranteed there.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The dataset webpage could be found with this DOI `10.17044/scilifelab.14687271`. All the instructions on how to access the data is clearly mentioned on the dataset landing page, which contains the actual data files along with metadata to help users better understand how to use the data.

When will the dataset be distributed?
The dataset has already been distributed with this DOI `10.17044/scilifelab.14687271`.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
Yes. Please visit the dataset home page for details about the license and terms.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
There is no third parties imposed IP-based or other restrictions on the data associated with the instances.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
There are no export controls or other regulatory restrictions on this dataset to the best of our knowledge.

Any other comments?
None.

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?
The data is currently supported/hosted by `https://scilifelab.figshare.com/` (the support letter could be seen on the final page of this article). The infrastructure for hosting and maintaining the data is guaranteed to be supported by the repository.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
The owners of the dataset could be contacted through either of the following email addresses: `yue3@kth.se` and `sorkhei@kth.se`.

Is there an erratum? If so, please provide a link or other access point.
There is no erratum for the dataset.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?
At the moment, there is no plan for any updates. In case the dataset is updated, the most recent version of it could be seen on the dataset website (previous versions will still be visible), and the DOI will also change accordingly with respect to the version.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
In case of retention, the data will be deleted and a new version that addresses the issue will be re-uploaded, in which case we ask users to delete their old copy of data (our ToU covers this). This is communicated clearly to the users.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
If there is a change in the version of the dataset, previous versions will still be hosted and supported on the website. We will announce the change of version as explicit as possible on the website.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
We always welcome if experts in the area of mammography are interested in contributing to our dataset by assessing mammographic potential masking of tumor in our mammograms. Code for our annotation tool with complete instructions on how to use it is publicly available at https://github.com/MoeinSorkhei/CSAW-M_Annotation_Tool/. For further discussion, experts are very welcome to contact us using the contact info on the website. We will then compare the received annotations against our ground truth using the same metrics we used in the paper. Finally, the annotations and the comparison against our ground-truth will be made publicly available on our website, acknowledging the contribution. We are also interested in receiving BI-RADS annotations. We would be happy to discuss any other possible contributions not mentioned here. We note, however, that although contributions will be made publicly visible on our website, they do not result in any change in the authors of the dataset.

Any other comments?
None.

Letter of support

I hereby confirm that the dataset “CSAW-M: An Ordinal Classification Dataset for Benchmarking Mammographic Masking of Cancer”, DOI: 10.17044/scilifelab.14687271, is currently hosted on the SciLifeLab Data Repository (https://scilifelab.figshare.com) and will receive continued support by the SciLifeLab Data Centre (https://scilifelab.se/data) for the foreseeable future.

The SciLifeLab Data Repository is a service for sharing research data offered to users of the research infrastructure SciLifeLab in Sweden (https://scilifelab.se). The SciLifeLab Data Repository is currently set up as an institutional instance of Figshare (based on a contract between Digital Science & Research Solutions Ltd and Uppsala University) maintained by the SciLifeLab Data Centre. The data is physically located at AWS servers in Europe as a subcontracted service under the Figshare agreement. We expect to transfer to an in-house storage solution and in the near future, but this will neither impact dataset availability nor the assigned DOI. In addition, we plan to build a number of services around the repository to allow easier usage of the deposited data for annotation and analyses. In particular, a service for sharing and deploying trained AI models to process submitted datasets will probably be of particular interest for users of the data published in association with the currently submitted manuscript.

At the time of writing this letter, the dataset “CSAW-M: An Ordinal Classification Dataset for Benchmarking Mammographic Masking of Cancer” is not by default available for anyone to download but requires a request for access with a motivation. The authors of the dataset themselves formulate the conditions for granting access, decide on approving or rejecting access requests, and send information about how it can be downloaded in case access is granted. The SciLifeLab Data Centre does not participate in this decision but only ensures that the data is stored safely and stays available to those who were granted access and received information about how to access it.

Johan Rung,
Head of SciLifeLab Data Centre
Dept. of Immunology, Genetics and Pathology, Uppsala university
Box 815
751 08 Uppsala
johan.rung@scilifelab.uu.se
tel. +46 (0) 72-2509211