This article has been accepted for publication in IEEE Transactions on Medical Imaging. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMI.2021.3129068

Multi-task fusion for improving mammography screening data classification

Maria Wimmer, Gert Sluiter, David Major, Dimitrios Lenis, Astrid Berg, Theresa Neubauer, and Katja Bühler

***Abstract*— Machine learning and deep learning methods have become essential for computer-assisted prediction in medicine, with a growing number of applications also in the field of mammography. Typically these algorithms are trained for a *specific task*, e.g., the classification of lesions or the prediction of a mammogram's pathology status. To obtain a comprehensive view of a patient, models which were all trained for the *same task(s)* are subsequently ensembled or combined. In this work, we propose a pipeline approach, where we first train a set of *individual, task-specific models* and subsequently investigate the fusion thereof, which is in contrast to the standard model ensembling strategy. We fuse model predictions and high-level features from deep learning models with *hybrid patient models* to build stronger predictors on patient level. To this end, we propose a multi-branch deep learning model which efficiently fuses features across different tasks and mammograms to obtain a comprehensive patient-level prediction. We train and evaluate our full pipeline on public mammography data, i.e., DDSM and its curated version CBIS-DDSM, and report an AUC score of 0.962 for predicting the presence of any lesion and 0.791 for predicting the presence of malignant lesions on patient level. Overall, our fusion approaches improve AUC scores significantly by up to 0.04 compared to standard model ensembling. Moreover, by providing not only global patient-level predictions but also task-specific model results that are related to radiological features, our pipeline aims to closely support the reading workflow of radiologists.**

***Index Terms*— Mammography, DDSM, CBIS-DDSM, Deep Learning, Model Fusion.**

I. Introduction

BREAST cancer is the most common cancer type in women and also the leading cause of death by cancer in women worldwide [1]. Fortunately, the mortality rate declined in recent years, one reason being the higher rate of early diagnosis due to the establishment of screening programs. Important cancer risk factors, such as breast density, can be detected and monitored early with such programs [1], [2].

Due to the increasing amount of imaging data, machine learning, especially deep learning algorithms are being developed to automatically process mammography data. Such models perform, for example, localization and classification of lesions [3], [4], breast density classification [5], [6], or cancer risk prediction [7], [8]. These automated methods can be used for accelerating reading workflows [9], [10], or ideally, to support radiologists in their image interpretation and diagnosis [11]. Several recent studies report higher accuracies when combining AI algorithms with the assessment of a single radiologist [12] or improved performance of radiologists when aided by an AI system [13], [14]. Besides the obtained performance gains, the assistance of radiologists as well as *human-computer collaboration* are becoming increasingly important aspects and challenges for future application in clinical practice [10], [15]. To increase trust in AI support tools, not only the interpretability of black box models is being intensively studied [16]–[18] but also the potential of providing intermediate model results that are linked to radiological features [9], [11]. Recent user studies in cancer screening and diagnosis showed that clinicians profited more from models that provide detailed results compared to solutions delivering solely a benign/malignant assessment [19], [20].

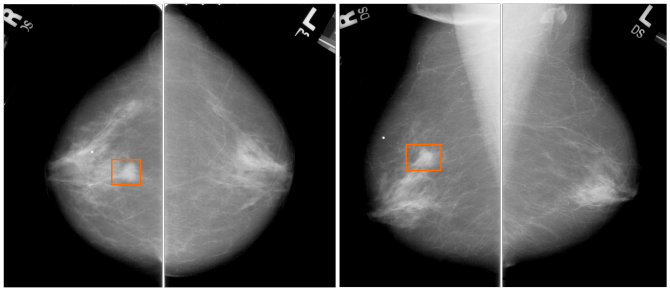


Fig. 1: Standard mammography study of a patient showing the four standard views (from left to right): R-CC, L-CC, R-MLO, and L-MLO. The patient has a malignant mass in the right breast, highlighted in orange. Best viewed in color.

A. Related Work

A standard mammography study (see Fig. 1) comprises four X-ray images that correspond to two different imaging views from each breast: L-CC, R-CC, L-MLO, and R-MLO. Thereby, CC corresponds to the craniocaudal (CC) view, MLO to the mediolateral oblique (MLO) view, and L and R indicate the left or right breast, respectively. Radiologists analyze each view in detail and compare them to obtain a comprehensive

The authors are with VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH, 1220 Vienna, Austria (Corresponding author: Maria Wimmer, mwimmer@vrvis.at).

VRVis is funded by BMK, BMDW, Styria, SFG, Tyrol, and Vienna Business Agency in the scope of COMET - Competence Centers for Excellent Technologies (879730) which is managed by FFG. Thanks go to our project partner AGFA HealthCare for valuable input.

view of a patient and render a diagnostic decision [21]. Suspicious lesions, for example, can be visible in one view of a breast but may be obscured in the other view. Therefore, a thorough analysis is necessary. Various deep learning-based methods have been presented in the past years that analyze single- or multiple-view images at a time. However, this is strongly dependent on their task and related clinical question.

1) Breast Density Scoring: Breast density is an important risk factor as dense breast tissue is related to the development of cancer. Furthermore, microcalcifications (MCs) and masses are harder to see on the mammograms, causing misdiagnoses [2]. The BI-RADS standard [22] defines density in four categories (a-d) as a measure of the breast tissue composition: "almost entirely fatty" (a), "scattered areas of fibroglandular density" (b), "heterogeneously dense" (c), and "extremely dense" (d). Assessment by radiologists usually has a high inter-observer variability [23] due to the qualitative description of the four categories. Therefore, automated density classification models often focus on the two superclasses *not dense* or *fatty* (a+b) and *dense* (c+d) [6].

Recent works utilized all four mammography views via multi-view CNNs to classify breast density into the four density categories [24] or in both superclasses [6], [24]. In contrast, Lehman et al. [5] trained a ResNet-18 model to classify single mammograms and assigned the consensus density across all views for the patient. Other methods used a refined AlexNet to classify the two middle density classes (b+c) [25] or performed unsupervised feature learning to segment dense tissue and derive a density scoring per image [26].

2) Lesion Localization and Classification: Exact localization and classification of lesions (i.e., masses, calcifications, and clusters of MCs) in mammograms are crucial as they are important risk factors or already indicators of cancer [21] (see Fig. 2). While many works perform lesion localization, quantification, classification, or all together [3], [4], [27]–[29], others solely classify already extracted lesions on patches [11], [30]–[33]. The use of classical feature extraction and machine learning methods, or the combination thereof with CNNs, has been intensively investigated in the literature [4], [27], [30], [34], [35]. Mordang et al. [32] were the first to use CNNs for MC localization and utilized a VGG-like architecture for this task. Various studies focused on the classification of MCs and MC clusters [33], [36], [37], e.g., with a combination of a difference-of-Gaussians detector and two-stream CNNs [36]. Dhungel et al. [38], among many others [11], [30], [31], [35], [39], performed localization and analysis of masses. They combined deep belief networks, Gaussian mixture models, and CNNs for mass detection. Barnett et al. [11] recently proposed an interpretable mass classification framework with the goal to follow the reasoning process of radiologists. Finally, state-of-the-art object detection approaches like Faster R-CNN [40] or YOLO have been applied for lesion localization and classification [3], [29], [39], [41], [42]. Ribli et al. [3] utilized a Faster R-CNN with a VGG16 backbone to detect and classify lesions into malignant and benign classes individually. Others extended a Faster R-CNN model by a cascaded classification step to reduce false-positive detected lesions [29].

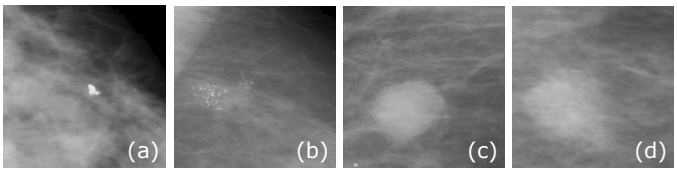


Fig. 2: Patches showing a benign calcification (a), malignant MC cluster (b), benign mass (c), and a malignant mass (d).

3) Malignancy Scoring: Several studies classify single or multiple mammograms directly to obtain a score assessing whether the view image is cancerous [7], [18], [42]–[44] or contains a (specific) malignant or benign finding [45]–[48]. Recent works utilized, e.g., an all-convolutional design combined with curriculum learning [43], multi-instance learning [18], [42], [45], [49], self-supervised methods [48], or a multi-view-multi-task approach [9]. Wu et al. [46] concatenated heatmaps obtained from sliding window patch classification to classify full images. Other works derive a malignancy score per view image, breast, or patient by averaging or considering the maximum score, e.g., obtained from a Faster R-CNN [3] or a map of pixelwise abnormality scores [13].

4) Feature or Information Fusion: The fusion of *features* or, more generally, of (extracted) *information* is inspired by how radiologists assess and compare ROIs and mammograms to obtain a comprehensive view of a patient. The term *feature* can refer to "classical, handcrafted" features (e.g., gabor filters, curvelets, entropy, etc.), CNN-features extracted by a CNN, or non-imaging features (e.g., patient age). The extraction and fusion can be performed at different scales, for example, *locally* from/within a single-view image, patches or across ROIs [4], [18], [33], [42], [44], [45], [49]. Kooi et al. [4] fused CNN and classical features extracted from patches within a single mammogram. Lotter et al. [44] and Shen et al. [18] fused local CNN patch features, whereas the former extracted them with a sliding window approach, and the latter extracted only CNN features from salient regions obtained with a global image classifier. Another common approach is to utilize *multiple views* for localization and classification of lesions and full images, as summarized by Jouirou et al. [50]. While Shachor et al. [37] dynamically combined classical features from local patches from MLO and CC view for calcification classification, Kooi et al. [27] fused CNN features from ROIs across views for malignant mass detection. The usage of *multi-view CNNs*, where each view image is processed with a CNN, followed by feature fusion at a given layer, has been studied as well [28], [51], [52] for different purposes, e.g., BI-RADS scoring [51] or breast density classification [6], [24]. McKinney et al. [7] developed several models, which used different fusion and combination strategies, e.g., concatenation of CNN patch features across all views, fusion of CNN image-level features per breast, and/or patient, or concatenation of non-imaging features (e.g., patient age) with CNN features.

The last stage is *decision-level fusion*, i.e., fusion of predictions, which has been investigated by Kyono et al. [52], for example. They predicted several radiological features (e.g., breast density, diagnosis, age) with a multi-task CNN sepa-

rately for each view, fused them, and classified the patient as benign or malignant. Finally, the naive ensembling of predictions from different models, e.g., via averaging, can also be interpreted as decision fusion [3], [7], [13].

5) Summary: While many recent works directly classify ROIs or view images with, e.g., CNNs, a significant part utilizes some form of information fusion when processing mammography data (see Table I). The reasons are manifold: fusion is performed to *(i)* incorporate different aspects at different levels (ROI, image, patient), *(ii)* thus, increase robustness and performance of classification models [4], [18], [27], [33], [37], [42], [44], [45], [49], and *(iii)* increase explainability and interpretability of model predictions [9], [11], [17], [18], [52]. Methods that fuse predictions across one or more ROIs or mammograms usually build upon models that predict the same scores for the same task or perform standard model ensembling strategies [5], [7], [9], [52]. On the other hand, methods that perform a fusion of features within or across images mostly do not provide intermediate results (e.g., assessment of suspicious regions) but only final classification results. Although recent user studies highlight the potential of providing detailed classification results or pinpointing to suspicious regions [19], [20], only a few proof-of-concept studies explored fusion and the potential of providing intermediate results similar to the assessment of radiologists in the field of mammography [9], [11], [52]. These methods operated only on lesion-level [11] or fused models that predict the same multi-task scores [9], [52]. To the best of our knowledge, the fusion of models trained for *different* tasks is not being studied in the context of mammography.

B. Contribution

Closing this gap, we investigate information fusion for mammography from another perspective by focusing on the fusion of features and predictions from *individual, task-specific models* to obtain a comprehensive assessment on patient level. To this end, we propose a *pipeline approach* comprising

- the development of three *task-specific models*, namely *(i)* a breast density classification model, *(ii)* a lesion localization model, *(iii)* and a findings classifier, as a basis for fusion, and
- the investigation of two fusion strategies: *(i)* the fusion of high-dimensional, task-specific CNN features with a *multi-input embedding CNN* and *(ii)* prediction score fusion of model predictions with MLPs.

By building upon task-specific features and decisions, we obtain *hybrid patient meta-models*, which access these intermediate results in their prediction. Due to the two-stage nature of our method, we report not only a global score on patient level but make the sub-results that reflect radiological features also accessible to the clinician.

We train both fusion approaches for two different classification targets, which we will refer to as *patient predictions* (i.e., the prediction of the respective model). We predict *(i)* the presence of *any* lesion (*lesion prediction*), *(ii)* and whether the patient has any *malignant* lesion (*malignancy prediction*).

At each stage in our pipeline, we aim for resource-efficient models, and therefore, utilize lightweight architectures like MobileNets [53] for image classification-related tasks. The full pipeline was trained and evaluated on the well-known and publicly available DDSM [54], [55] and CBIS-DDSM datasets [56], [57]. In a comprehensive technical analysis, we show that our task fusion strategy improves patient-level classification over standard model ensembling. A detailed analysis of results and discussion thereof as well as future clinical perspectives are provided in the discussion.

II. Materials and Methods

We define a set of mammography images $\mathcal{I}_i = \{I_i^v\}$ for patient i and mammography image view $v \in \{\text{L-CC, L-MLO, R-CC, R-MLO}\}$. We will refer to this set $\mathcal{I}_i$ as *exam* or *case* of patient i.

A. Data

We utilize two publicly available mammography databases for our experiments: the Digital Database for Screening Mammography (DDSM) [54], [55] and its curated version CBIS-DDSM [56], [57].

1) DDSM and CBIS-DDSM Dataset: The *original DDSM dataset* [54], [55] comprises 2620 mammography screening exams $\mathcal{I}_i$, collected from four different sites acquired with four different scanners. The data is grouped in four categories:

- *normal* (695 cases): normal exams with no suspicious abnormalities and proven normal exams four years later
- *benign without callback* (141 cases): cases with benign abnormality but without need for callback
- *benign* (870 cases): including suspicious findings which were identified as benign findings after callback
- *cancer* (914 cases): cancer was proven via histology

An expert radiologist labeled the breast density per patient and provided pixel-level annotation for abnormalities. Each abnormality is described following the BI-RADS standard [22], including lesion type (mass or calcification) and further details like shape, lesion margin, and calcification type.

The *CBIS-DDSM dataset* [56], [57] was published at The Cancer Imaging Archive [58] as curated version of the original DDSM set, whereby only images showing one or more lesions have been transferred. Annotated masses were re-checked by a radiologist, and pixel-wise annotations have been refined with an automated segmentation algorithm. However, annotations of calcifications remained unchanged. The authors also provided a predefined split into train and test sets to ensure comparability between methods evaluated on this dataset. Overall, the CBIS-DDSM dataset comprises 3568 annotated lesions (1696 masses, 1872 calcifications) in a total of 3032 mammography view images. For further details on the data, we refer to the original publications [56], [57].

2) Data Harmonization and Preparation: While providing enhanced annotation quality, the CBIS-DDSM dataset has two shortcomings: first, the absence of normal images without lesions, and second, the lack of full patient mammography exams including all four views. To utilize both resources

TABLE I: Overview on related works. Target: density = breast density classification, lesions = lesion localization and/or classification, malignancy = prediction of BI-RADS, benign/malignant, cancer yes/no, etc., on image/patient level; Data: name of image database; Fusion: ✓= some form of fusion involved; Intermediate / Sub-results: type of intermediate/additional results provided apart from final scores; Method: brief summary (loc. = localization, seg. = segmentation, class. = classification, RF = random forest, DBN = deep belief network, GMM = gaussian mixture model, DoG = difference of gaussian).

Author	Target	Data	Fusion	Intermediate / Sub-results	Method
[6], [24]	density	private	✓	no	multi-view CNN
[5], [25]	density	private		no	single-view CNN
[26]	density	private	✓	dense tissue segmentation	multi-scale unsupervised seg. + texture scoring
[38]	lesions	INbreast		no	DBN + GMM (loc.), CNN + RF (class.)
[34][1], [30][2]	lesions	BCDR[2], private[1]		no	SVM[1] / CNN + SVM[2] for class.
[31][1], [32][2], [35][3]	lesions	DDSM[1], private[1,2], CBIS-DDSM[3], INbreast[3]		no	CNN for class.[1,3] / loc. + class.[2]
[36]	lesions	private	✓	no	DoG + multi-scale two-stream CNN
[37]	lesions	DDSM	✓	no	multi-view CNN
[33]	lesions	DDSM	✓	no	classical features + feed forward network
[4]	lesions	private	✓	no	candidate loc. (RF) + class. (CNN features + classical texture features)
[27]	lesions	private	✓	no	dual-stream CNN for lesion ROI class.
[11]	lesions	private	✓	class activation map, mass margin class score	case-based reasoning, compares parts of new images to learned prototypes
[3][1], [41][2] [29][3], [39][4]	lesions	DDSM[1,4], INbreast[1,2,3], OPTIMAM[2], private[1,3]		no	Faster R-CNN/YOLO-based lesion localization + classification
[42]	lesions, malignancy	DDSM, OPTIMAM, private	✓	benign + malignant lesions (bounding boxes)	RetinaNet-based approach + multi-stage training (fully + weakly supervised, multi-instance learning)
[48]	lesions, malignancy	INbreast, private	✓	malignancy probability map	self- and weakly supervised reconstruction for lesion loc./seg., image-level class.
[28]	malignancy	INbreast, DDSM	✓	no	multi-view CNN
[43]	malignancy	CBIS-DDSM, INbreast		salient regions	all-convolutional CNN (two-stage)
[45]	malignancy	INbreast	✓	no	multi-instance approach
[46], [51]	malignancy	private	✓	heatmaps of malignant / benign+malignant regions	multi-view CNN(s), fusion at different stages
[47]	malignancy	INbreast, CBIS-DDSM		malignant regions	CNN + region-based/global group-max pooling
[13]	malignancy	OPTIMAM, private		pixel-wise abnormality score	semi-supervised CNN (two-stage)
[44]	malignancy	DDSM	✓	no	multi-scale CNN + curriculum learning
[18]	malignancy	private	✓	saliency maps (malignant findings)	weakly supervised approach, global (weak loc.) + local CNN
[9], [52]	malignancy	private	✓	radiological features per view, heatmaps	multi-view, multi-task CNN
[7]	cancer risk	OPTIMAM, CBIS-DDSM, private	✓	malignant regions (bounding boxes)	patch-level, image-level and CNN+non-imaging feature fusion (various models)
This work	density, lesions, malignancy	DDSM, CBIS-DDSM	✓	breast density, lesions (bounding box + label), findings classification	task-specific CNNs (multiple scales), feature and prediction fusion with CNNs + MLPs

without losing their individual benefits, we prepare the data as follows:

First, we preprocess the DDSM set in the same way as it was done for the CBIS-DDSM data, including optical density normalization and remapping the data to the full 16-bit range [1].

Next, we match, i.e., compare the CBIS-DDSM images to the preprocessed DDSM data to identify corresponding cases and obtain a total of 2590 full mammography exams. We assign the malignancy status of a lesion according to the curated annotation from CBIS-DDSM, whereby "benign without callback" will be treated as a benign case.

Finally, we identify potential ambiguous cases which have been originally in the cancer, benign, or benign without callback subset in DDSM but have not been transferred to CBIS-DDSM. Since the status of the lesions for these 329 cases remains unclear, we exclude them. Further, we exclude seven additional exams, which are either incomplete, i.e., not all four views are present, or appeared with different imaging data and annotations in different subsets of DDSM and CBIS-DDSM. This leads to our final set comprising *2254 cases*.

3) Train, Validation, Test Split: We split the dataset into train, validation, and test data on case-level and, thus, ensure that images from one case are not distributed across different sets. We preserve the train/test split of the data provided with the CBIS-DDSM set. The remaining normal cases are randomly distributed in the same ratio (∼80% training images) to the train/test set in a way that the distribution of breast density is similar in the three sets. From the obtained train set, we randomly select ∼12% of cases for the validation set in a way that the ratio of different breast density classes, lesion types, and pathology is similar across the three sets (see Table II). Overall, the train, validation, and test set comprise 1511, 290, and 453 cases, respectively. Out of the 2254 cases, 174 contain more than one lesion, with the maximum number of lesions per case being 24.

[1] https://github.com/fjeg/ddsm_tools

TABLE II: Distribution of breast density, lesion type, and pathology status in train, validation, and test set.

		Train	Validation	Test	Total
Density	a	207	40	50	297
	b	567	108	176	851
	c	448	86	134	668
	d	289	56	93	438
Lesion	normal	481	107	105	693
	mass	583	93	201	877
	calcification	485	96	150	731
Pathology	normal	481	107	105	693
	benign	522	98	199	819
	malignant	508	85	149	742

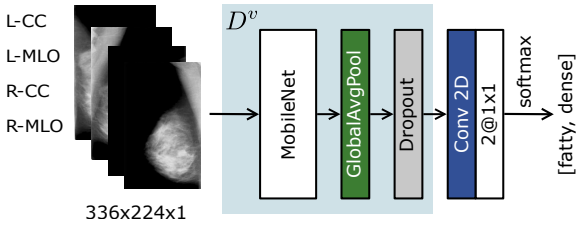


Fig. 3: Density view model D^v for view $v \in \{$L-CC, L-MLO, R-CC, R-MLO$\}$

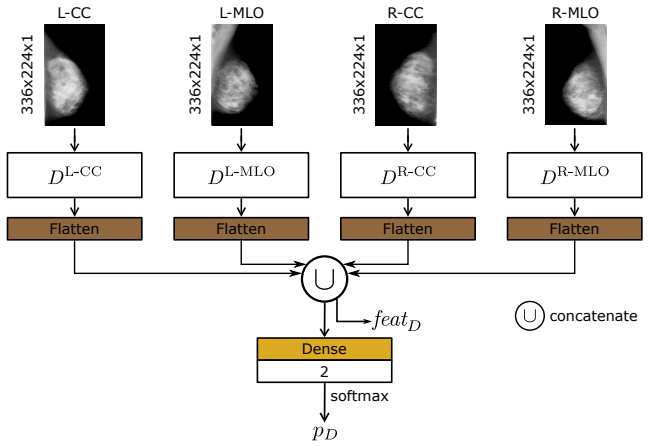


Fig. 4: Density patient model D

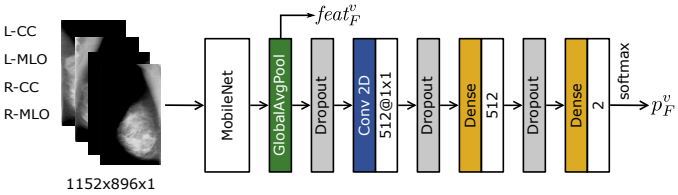


Fig. 5: Findings model F

B. Task-specific Mammography Models

The first stage in our pipeline is the development of a set $\mathcal{M}$ of three resource-efficient, task-specific models $\mathcal{M} = \{D, L, F\}$, which are the base for our patient model P:

- D performs breast density classification,
- L delivers bounding boxes around localized lesions and their respective class label, and
- F predicts the presence/absence of lesions in an image.

1) Breast Density Model (D): Radiologists include all four view images $\mathcal{I}_i$ in the assessment of a patient's breast density. Recent deep learning based density classification models follow this standard and utilize all views as input [6], [24], whereas the usage of only one view has also been studied [5]. We propose a two-stage approach where we employ both ideas in the design of density model D to increase robustness and classification performance.

We build a view model D^v first, which uses any single mammography image I_i^v as input to predict the density superclass, i.e., *fatty* or *dense*. The model is built upon a MobileNet classifier [53] with global average pooling, followed by a 1x1 convolution layer (see Fig. 3). Our final model D takes the four standard mammography views $\mathcal{I}_i$ as input where each image is passed to a separate branch (see Fig. 4). Each view branch consists of a density view model D^v, whereby the dropout rate is increased from 0.001 in model D^v to 0.5 in D. After the following flattening operation, the 1D feature vectors are concatenated, and a final dense layer predicts the density superclass. The obtained density score p_D at patient-level depicts the score corresponding to the "dense" class.

2) Findings Model (F): The objective of this model is to classify any single-view image I_i^v into "normal" or "image containing *any* findings", i.e., lesions. Such a model could be, for example, integrated into a reporting system, in which images with lesions are examined first by a medical expert. Again, we aim for a resource-efficient model to solve this task, and thus, we extend on our previous work [59], [60], where we already successfully applied MobileNet [53] in this context. Fig. 5 illustrates our findings model F with a MobileNet feature extractor and a modified classifier on top. Adding an additional dense and dropout layer increased the classification accuracy and the generalization capability of the model. Additionally, we use an increased dropout rate of 0.5 to stronger regularize the network. The output for each view image I_i^v is the score p_F^v which determines whether there is any lesion in I_i^v.

3) Localization Model (L): Similar to radiologists, we aim to detect the exact location of lesions within an image I_i^v and classify them into their correct type and malignancy status. The localization and characterization of lesions are important tasks, as they can be risk factors or already indicators of cancer [21]. Therefore, we develop model L to localize lesions and classify them in either "benign calcification", "malignant calcification", "benign mass", or "malignant mass". Inspired by recent works on lesion localization [3], [29], [41], we utilize the well-known Faster R-CNN [40] architecture. InceptionV2 [61] serves as feature extractor, which was already successfully applied in the context of mammography lesion localization [41]. Fig. 6 illustrates the architecture. Our localization model L classifies localized lesions into four types (benign calcification, malignant calcification, benign mass, and malignant mass) and assigns $k \in [0, n]$ scores $p_L^{v,k}$, depending on the number of detected lesions that are found in I_i^v.

C. Patient Meta-Model (P)

The hybrid patient meta-model P aims to efficiently combine the task-specific building blocks $\mathcal{M}$ to obtain a compre-

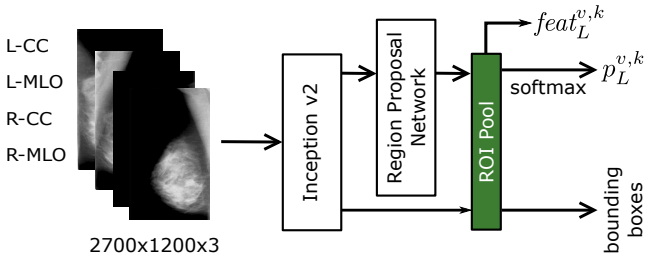


Fig. 6: Localization model L

hensive patient-level assessment while preserving the individual model predictions related to radiological features and risk factors. We consider two different patient predictions:

- *lesion prediction:* whether the patient has any lesion, regardless of pathology,
- *malignancy prediction:* whether the patient is malignant, i.e., has any malignant lesion.

The fusion of different models can be performed at various stages, whereby, again, our goal is to develop *resource-efficient* variants. For this, we compare the fusion of *prediction scores* as well as the fusion of *features* from the individual models.

1) Fusion of Predictions (P_{score}): The three task-models $\mathcal{M}$ deliver different prediction scores $p_m \in [0,1], m \in \mathcal{M}$ at various levels, i.e., patient level, image level, ROI level. We concatenate these predictions of the models introduced in Sec. II-B to form the vector w_p, formally:

$$w_p = p_D \cup p_F^v \cup p_L^{v,n} \tag{1}$$

where n is the number of considered detections per view. In case of no detected lesions by model L or less lesions than specified by n are found, a probability of 0 is assigned, indicating that no (additional) lesions have been localized. For the malignancy prediction, only scores $p_L^{j,n}$ corresponding to malignant masses and calcifications are considered in the combined scores vector w_p. In case no malignant lesions or less malignant lesions than specified by n are found, a value of 0 is assigned.

2) Fusion of Features (P_{feat}): Apart from the fusion of prediction scores p_m, we also propose the fusion of feature vectors $feat_m, m \in \mathcal{M}$ from the three different models. We extract features at the following stage in the networks:

- $feat_D$ is the 4096-dim., flattened, concatenated view-representations after global average pooling (see Fig. 4)
- $feat_F^v$ is the 1024-dim. representation for view image I_i^v, obtained after global average pooling (see Fig. 5)
- $feat_L^{v,k}$ is the 1024-dim. representation for detection k in I_i^v (see Fig. 6).

We propose an embedding network that takes the extracted, high-dimensional feature representations $feat_m$ as input in separate branches (see Fig. 7). Each channel corresponds to the respective features of a view image I_i^v. The density and findings branches consist of two convolution blocks, followed by pooling operations. The localization feature branch utilizes an additional convolution and pooling block for better feature

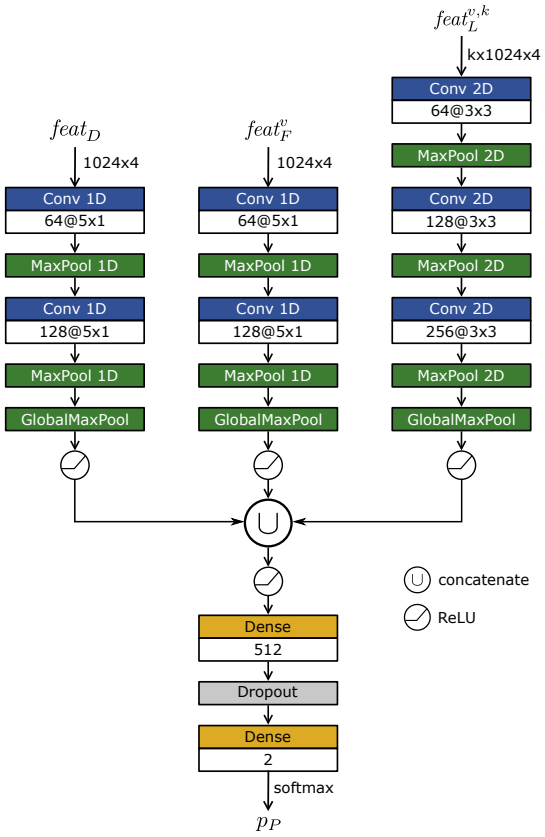


Fig. 7: Patient meta-model P_{feat}

learning. Before and after concatenation of all feature representations, we perform ReLU activations. The final classification part of the network consists of two dense layers with an intermediate dropout layer (dropout rate of 0.1) followed by a final softmax activation.

Again, we vary the number of lesions considered per view $n \in \{1,2,3,4,5\}$. In case no lesions are detected with model L, or less lesions than specified by n, background features are pooled from the feature map and used as input. For the malignancy prediction, only features $feat_L^{j,n}$ corresponding to malignant masses and calcifications according to the localization model L are considered for the feature fusion. In case of no malignant lesions or less than specified by n, again background features are considered as model input.

III. Experimental Setup

We implemented our framework in Python, utilizing Keras [62] with Tensorflow backend [63] for training the task-specific models D, D^v, L, and patient meta-model P_{feat}. Additionally, we used the Tensorflow Object Detection API [64] to train localization model L and scikit-learn for training patient meta-model P_{score}. Model training and experiments were conducted on an NVIDIA Titan X GPU (12 GB RAM).

A. Training Details

For the training of every task-specific model, we first segmented the breast with a basic, non-learning-based segmen-

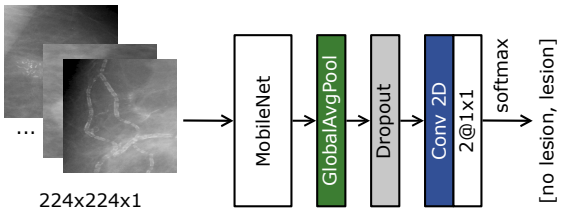


Fig. 8: Patch Model

tation approach according to Shen et al. [43][2]. Segmentation of the breast has been frequently used by related works as first preprocessing step, e.g., to clean/remove the background or for subsequent cropping to the breast area [43], [45], [47], [48]. Similarly, we used the obtained breast mask to clean the background and for the sampling of patches inside the breast for pre-training the findings model F (see Sec. III-A.2). The following set of random data augmentations was executed in each model training: horizontal flips, rotations (range: [-15,+15] degrees), and random sized crops (range: [85%,100%] of the image size). All image resizing operations were performed using bicubic resampling.

1) Breast Density Model: All images were resized to $336 \times 224 \times 1$ with rescaled intensities to the range [0,255] in floating-point precision to preserve the bit depth. Model training was conducted in a two-stage approach with Adam optimizer and cross-entropy loss: First, imagewise pre-training of the view model D^v (see Fig. 3) was performed for 25 epochs and an initial learning rate (lr) of 1e-3. Further, we employed Stochastic Weight Averaging (SWA) [65] with an initial epoch of 10 to increase the generalization capability of the model. In addition to the standard set of augmentations, random shears were conducted. Second, we trained the patient-wise model as shown in Fig. 4. Each view branch was initialized with the SWA-weights from stage 1, and the complete model was trained for 25 epochs (lr = 1e-4). SWA was used with an initial epoch of 5. Horizontal flipping was not performed to preserve the original position of the breast in each view but instead blurring and grid distortion were additionally carried out. We reduced the learning rate by a factor of 0.2 with a patience of 5 epochs on the validation loss in both training stages.

2) Findings Model: We performed two-stage training of the findings model F, a strategy already successfully applied by recent works [13], [42], [43]. In both stages, the models were optimized using Adam with cross-entropy loss. First, a patch classifier (see Fig. 8) was trained from scratch with patches of size $224 \times 224 \times 1$, inspired by Shen et al. [43]. We extracted an initial set by sampling 5 patches per lesion (overlap $>$ 90% with lesion) and 5 patches from normal images (overlap $>$ 90% with breast). The patch model was trained with a batch size of 64 (lr = 1e-4) and early stopping on the validation loss (patience = 10 epochs, tolerance = 0.001). Additional augmentations were performed (vertical flips, transpose, and shift/scale/rotate) to further increase the diversity of patches. The model was fine-tuned in a second training iteration with a reduced learning rate of 1e-5.

[2]https://github.com/lishen/end2end-all-conv

In the second stage, we initialized the feature extractor of the findings model F (see Fig. 5) with the obtained patch weights. The full images were resized to $1152 \times 896 \times 1$, rescaled to [0,1] and z-score normalized. F was trained using a batch size of 6 (lr = 1e-4). As opposed to the patch model, the validation AUC score was monitored as the criterion for early stopping (patience = 10 epochs, tolerance = 0.001). Additionally, SWA was used with an initial epoch of 5 which further improved the generalization capability. The model was fine-tuned in a second training round with lr = 1e-5. In both training iterations of model F, vertical flips were additionally performed. In addition, stratified sampling was used to balance batches between images showing lesions and normal images [66].

3) Localization Model: The InceptionV2 backend was initialized with COCO-weights and then fine-tuned for the mammography lesion localization task for the four classes. The ground truth bounding boxes required to train the Faster R-CNN model were derived from the pixelwise annotated lesions. We consider the axis-aligned minimum bounding box which encloses the lesion. The model was trained according to the pipeline split, whereby only images with at least one lesion were considered for training. We resized the view images to 2700×1200 and trained L with SGD (momentum = 0.9, lr = 1e-4) for 100k iterations and a batch size of 2. In addition to the default data augmentation strategies, bounding boxes were randomly jittered with a ratio of 0.005.

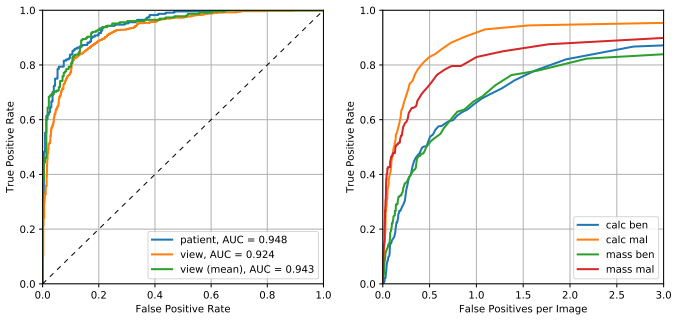
4) Patient Meta-Model: We performed a parameter search over the number of considered lesions $n \in \{1, 2, 3, 4, 5\}$ for P_{score} and P_{feat} and trained all models according to the predefined data split for the lesion and malignancy prediction. Best models were selected based on validation AUC and recall.

a) Prediction score fusion: Prediction scores were concatenated according to Eq. II-C.1 to obtain one feature vector w_p per patient. We varied the number of detected lesions $n \in \{1, 2, 3, 4, 5\}$ considered per view and included only their scores. For comparison, a classic SVM with RBF kernel, a multilayer perceptron (MLP), and a random forest were trained. Parameter search was performed over the parameters of the individual models and selected the model with highest validation AUC: SVM RBF (C = {1e-1, 1e-2, 1e-3, 1e-4, 1, 10, 100, 500, 1000}), random forest (number of trees = {3,5,7,10,15,20}), MLP (layer configuration = {[$|w_p|$, 2], [$|w_p|$, $|w_p|$, 2], [$|w_p|$, $|w_p|$/2, 2]}) .

b) Feature fusion: Before feeding the feature representations to P_{feat}, they were normalized with ϕ, where $\phi : \mathbb{R}^n \mapsto [-1, 1]$, resulting in normalized representations $\phi(feat_D)$, $\phi(feat_F^j)$, and $\phi(feat_L^{j,k})$. We optimized P_{feat} with Adam using cross-entropy loss and a batch size of 8 and lr = 5e-4. Early stopping was used with a patience of 10 epochs on the validation loss (tolerance = 0.001). Again, batches were balanced to ensure equal distribution of classes.

B. Evaluation Metrics

We compare the performance of classification-related tasks D and F by calculating widely used metrics in the field: the true positive rate (TPR), also referred to as *sensitivity* or *recall*, the true negative rate (TNR), also referred to as *specificity*, accuracy, and F1-score (F1), i.e., the harmonic mean of precision



(a) breast density D (b) lesion localization L

Fig. 9: ROC and FROC curves of individual models.

and recall. Further, we calculate the area under the ROC curve (AUC), which shows the TPR against the false-positive rate (1 - specificity). Additionally, we provide the area under the precision-recall curve (AUPRC) for comparisons with recent studies [18], [37], [52]. For the localization model L, we provide FROC curves to measure its detection performance and calculate the number of false positives per image (FPI) at given TPR rates.

IV. Results and Discussion

This section summarizes intermediate results obtained with task-specific models (see Section IV-A) as well as final predictions obtained via score and feature fusion (see Section IV-B). Section IV-C summarizes ablation study results, and finally, Section IV-D provides an in-depth discussion and analysis of the presented results.

We performed Wilcoxon signed-rank tests on the predictions for task-specific models, fusion models, as well as for the performed ablation studies. Similar to recent studies [14], [47], [48], we set the significance level to $\alpha = 0.05$.

A. Performance on Individual Tasks

Task-specific model results are visualized in terms of ROC and FROC curves in Fig. 9.

1) Breast Density Classification: We report an AUC score of 0.948 of density model D on the test set with TPR = 0.882 and specificity = 0.832 (F1 = 0.861). As depicted in Fig. 9a, the final model D on patient level (blue) shows a minor improvement in terms of AUC (AUC = 0.943, $p < 0.001$) compared to the aggregated predictions mean(D^v) of the imagewise model $D^v, v \in \{\text{L-CC, L-MLO, R-CC, R-MLO}\}$ on patient level (TPR = 0.833, specificity = 0.889, F1 = 0.858). Further, we observe a significantly higher sensitivity with D^v compared to mean(D^v) ($p < 0.001$) at similar accuracies (see Table III). On image level, we report an AUC of 0.924 with D^v (TPR = 0.815, specificity = 0.894, F1 = 0.849, accuracy = 0.854).

Table III summarizes density classification results reported in the literature. We report higher accuracy scores on DDSM compared to Oliver et al. [67], who tested only on a subset of 831 R-MLO images, while our method was evaluated on 453 patients, i.e., 1812 view images. While our model performs slightly beneath published works, these methods were trained utilizing significantly larger datasets, e.g., the dataset by Wu et al. [24] comprises 200k exams (80% train / 20% test data).

TABLE III: Overview on reported density classification accuracies (acc.) in related works and obtained with our model D. Methods indicated with * use one image as input, those without utilize all four view images.

Method	Data	Acc. (4 cls)	Acc. (2 cls)
Wu [24]	private (NYU)	0.767	0.865 (derived)
Lehman [5] *	private	0.770	0.870 (derived)
Kaiser [6]	private	–	0.881
Oliver [67] *	DDSM (R-MLO)	0.772	0.842
Ours * (D^v)	DDSM	–	0.854
Ours (mean(D^v))	DDSM	–	0.861
Ours (D)	DDSM	–	0.857

2) Findings Classification: For the task of classifying images into those with any lesion and those without, model F reaches an AUC score of 0.921 on test data with TPR = 0.881 and specificity = 0.802 (F1 = 0.878).

To the best of our knowledge, there is only the work by Lotter et al. [44], who used the presence/absence of lesions as classification target for pre-training their model on patch-level, thus, they did not report performance measures on image level.

3) Lesion Localization: We report TPR rates of 0.84 for malignant masses, 0.93 for malignant calcifications, 0.70 for benign masses, and 0.68 for benign calcifications by localization model L on test images with lesions, as summarized in Table IV. Fig. 9b shows corresponding FROC curves. A lesion is considered detected if the intersection over union (IoU) of the detected bounding box with the ground truth bounding box is ≥ 0.2, or if the center of the detected bounding box lies within the ground truth bounding box [3]. On normal images in the test set (105 patients, i.e., 420 view images), we detect 386 false-positive lesions in 188/420 images. On the 348 abnormal cases, we detect 2478 false-positive lesions.

Fig. 10 shows visual samples of correct and false-positive detected lesions. Overall, we report lower detection rates for benign lesions compared to malignant lesions, a phenomenon also observed in the literature [41]. As visible in Fig. 10, one reason for the lower performance of model L is the detection of small calcifications (in blue), which appear very similar to benign calcifications but are not annotated as such in the ground truth. Another aspect is the misclassification of denser breast tissue with masses as well as overlaps of benign and malignant masses that can occur due to non-maxima suppression performed on class-level.

Table IV provides an overview on localization results reported in the literature. However, the localization performances of the different methods cannot be compared directly due to the large differences in the datasets and varying criteria for correctly detected lesions. The method by Agarwal et al. [41], for example, utilizes the much larger OPTIMAM database, while Anitha et al. [68], on the other hand, use only a subset of the DDSM set rather than the full dataset.

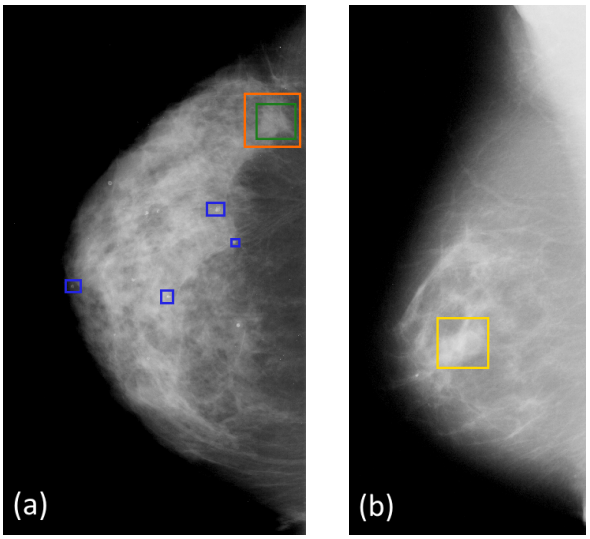


Fig. 10: (a) R-CC image with correctly localized malignant mass (green = ground truth, orange = detected) and additional detected benign calcifications (blue) not present in ground truth, (b) R-MLO image with false-positive benign mass (yellow). Best viewed in color.

TABLE IV: Overview on lesion localization results reported in related works and results obtained with our model *L* (OMI-H = OPTIMAM database, Hologic scanner images only, * = subset of 300 images used).

Method	Train/Test Data	Lesion	TPR @ FPI
Agarwal [41]	OMI-H / OMI-H OMI-H / INbreast	mass mal. mass ben. mass	0.93 @ 0.78 0.99 @ 1.17 0.85 @ 1.0
Ribli [3]	DDSM, private / INbreast	mal. lesion	0.9 @ 0.3
Akselrod-B. [29]	private / INBreast, private	mass	0.90 @ 0.3
Anitha [68]	– / DDSM*	mass	0.925 @ 1.06
Ours	DDSM / DDSM	mal. mass mal. calc. ben. mass ben. calc.	0.84 @ 1.0 0.93 @ 1.09 0.70 @ 1.06 0.68 @ 1.06

B. Patient Meta-Model Results

ROC curves for feature and score fusion P_{feat} and P_{score} (with MLPs), respectively, for both patient predictions, are shown in Fig. 11. Table V summarizes quantitative performance measures on test data. We additionally trained our fusion models without density information (indicated with * in Table V) and compared all our fusion results to standard ensembling, i.e., taking the maximum of prediction scores. A detailed statistical significance analysis for all fusion models is summarized in Table VII.

For P_{score}, we obtain the best results in terms of AUC and TPR with MLPs for both patient predictions, compared to SVMs and random forests (see Table VI). In terms of the number of included lesions n in the meta-models, the best results reported in Table V and Table VI are obtained with $n = 3$ for the lesion prediction (P_{score} and P_{feat}), and $n = 3$ (P_{score}) and $n = 1$ (P_{feat}) for the malignancy prediction. A detailed overview of quantitative results for P_{score}

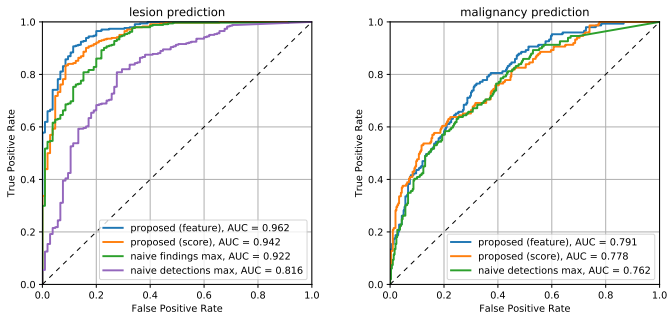


Fig. 11: ROC curves of patient models P_{score} and P_{feat} for the lesion prediction (left), and malignancy prediction (right).

TABLE V: Performance metrics of patient fusion models P_{score} (MLPs) and P_{feat} on test data. Models marked with * indicate exclusion of breast density information. $\max(p_F^v)$ and $\max(p_L^v)$ denote the naive score maximum of findings model *F* and detection model *L* for lesion prediction and malignancy prediction, respectively.

Model	Target	AUC	F1	TPR	Specificity
P_{score}	lesion	0.942	0.932	0.933	0.771
P_{score}*	lesion	0.941	0.928	0.919	0.800
$\max(p_F^v)$	lesion	0.922	0.938	0.974	0.667
P_{feat}	lesion	0.962	0.948	0.956	0.800
P_{feat}*	lesion	0.959	0.943	0.939	0.829
$\max(p_F^v)$	lesion	0.922	0.938	0.974	0.667
P_{score}	malignancy	0.778	0.601	0.591	0.813
P_{score}*	malignancy	0.774	0.523	0.578	0.857
$\max(p_L^v)$	malignancy	0.762	0.578	0.570	0.800
P_{feat}	malignancy	0.791	0.603	0.638	0.763
P_{feat}*	malignancy	0.789	0.581	0.577	0.797
$\max(p_L^v)$	malignancy	0.762	0.578	0.570	0.800

(SVM, random forest, MLP) and P_{feat} for different numbers of included lesions n is provided in the supplemental material.

Overall, we report an increase in terms of AUC between 0.02 and 0.04 for the lesion prediction with score fusion and feature fusion, respectively, when comparing to the naive score maximum across the four views $\max(p_F^v)$ ($p < 0.001$ for both). A slightly smaller increase is obtained for the malignancy prediction, ranging from 0.016 (P_{score}) to 0.029 (P_{feat}), as compared to $\max(p_L^v)$ ($p < 0.001$ for both). We report higher AUC scores and increased sensitivity, i.e., TPR, with feature fusion models compared to score fusion models for both patient predictions ($p < 0.001$). However, for the malignancy prediction, we observe a reduced specificity for feature fusion as compared to score fusion.

TABLE VI: Comparison of performance metrics of P_{score} for MLP, SVM, and random forest for both patient predictions.

Model	Target	AUC	F1	TPR	Specificity
MLP	lesion	0.942	0.932	0.933	0.771
SVM	lesion	0.935	0.928	0.916	0.810
Random Forest	lesion	0.929	0.924	0.919	0.771
MLP	malignancy	0.778	0.601	0.591	0.813
SVM	malignancy	0.763	0.552	0.483	0.867
Random Forest	malignancy	0.776	0.581	0.564	0.813

TABLE VII: Statistical significance analysis for fusion models P_{score} and P_{feat} with p-values $p < 0.05$ denoting statistically significance (bold font). Models marked with * indicate exclusion of breast density information. $\max(p_F^v)$ and $\max(p_L^v)$ denote the score maximum of findings model F and detection model L for lesion and malignancy prediction, respectively. p-values are given for the lesion prediction and the malignancy prediction (separated by "/").

Model	P_{score}	P_{score}*	P_{feat}	P_{feat}*	$\max(p_F^v)$	$\max(p_L^v)$
P_{score}	∞	$<$ **0.001** / $<$ **0.001**	$<$ **0.001** / **0.023**	$<$ **0.001** / **0.002**	$<$ **0.001** / —	$<$ **0.001** / $<$ **0.001**
P_{score}*	$<$ **0.001** / $<$ **0.001**	∞	$<$ **0.001** / $<$ **0.001**	$<$ **0.001** / 0.237	$<$ **0.001** / —	$<$ **0.001** / 0.999
P_{feat}	$<$ **0.001** / **0.023**	$<$ **0.001** / $<$ **0.001**	∞	$<$ **0.001** / $<$ **0.001**	$<$ **0.001** / —	$<$ **0.001** / $<$ **0.001**
P_{feat}*	$<$ **0.001** / **0.002**	$<$ **0.001** / 0.237	$<$ **0.001** / $<$ **0.001**	∞	$<$ **0.001** / —	$<$ **0.001** / 0.920
$\max(p_F^v)$	$<$ **0.001** / —	$<$ **0.001** / —	$<$ **0.001** / —	$<$ **0.001** / —	∞	$<$ **0.001** / —
$\max(p_L^v)$	$<$ **0.001** / $<$ **0.001**	$<$ **0.001** / 0.999	$<$ **0.001** / $<$ **0.001**	$<$ **0.001** / 0.920	$<$ **0.001** / —	∞

Fig. 12 shows a sample result obtained with our proposed mammography pipeline: The localization model L was able to correctly localize the malignant mass in the right breast but also falsely detected a benign mass at the same location. In R-MLO view, the benign detection was given a higher confidence than the malignant detection, which would lead to false patient-level results in case we solely rely on L. However, the feature fusion models P_{feat} were able to correctly classify the patient in terms of the lesion and malignancy prediction, while P_{score} failed for the malignancy prediction (decision threshold = 0.5).

C. Ablation Studies

Complementary to the training setup described in Section III-A, we performed additional experiments to support our pre-training strategies for task-specific models D and F. Further, we investigated the influence of breast density information in the fusion models.

1) Pre-training of D: We retrained the density model D without pre-training the view model D^v with the same training parameters (see Section III-A.1) except for a lower learning rate of 1e-3. We obtain a significantly lower AUC score of 0.900 ($p < 0.001$) with this model. Further we report TPR = 0.934, specificity = 0.690, F1 = 0.817, and a significantly lower accuracy of 0.797 as compared to D ($p < 0.001$).

2) Pre-training of F: Further, we retrained the findings classifier F without patch-wise pre-training with the same training parameters as described in Section III-A.2. The model without pre-training achieves a significantly lower AUC score of 0.895 ($p < 0.001$) and sensitivity = 0.816, F1 = 0.846, specificity = 0.817.

3) Breast Density Ablation Study: As breast density is an essential risk factor for breast cancer [2], we retrained our patient meta-models with the same training parameters but excluded breast density features and scores for P_{feat} and P_{score}, and denoted the obtained models P_{feat}* and P_{score}*, respectively. Results in Table V show higher AUC scores and a higher TPR for all fusion models P_{score} and P_{feat} when *including* breast density information ($p < 0.001$, as summarized in Table VII). No statistically significant difference can be reported when comparing P_{score}* and P_{feat}* with $\max(p_L^v)$ for the malignancy prediction with p-values $p = 0.999$ and $p = 0.920$, respectively. Further, no significant difference can be observed between P_{score}* and P_{feat}* for the malignancy prediction ($p = 0.237$). These results indicate that the inclusion of breast density can yield improved classification performance.

TABLE VIII: Comparison of different measures obtained with D and $\text{mean}(D^v)$ at various decision thresholds.

Model	Threshold	TPR	F1	Spec.	Acc. (2 cls)
D	0.6	0.860	0.871	0.885	0.872
$\text{mean}(D^v)$	0.6	0.781	0.838	0.916	0.848
D	0.7	0.816	0.859	0.916	0.866
$\text{mean}(D^v)$	0.7	0.711	0.804	0.942	0.826
D	0.8	0.763	0.841	0.947	0.855
$\text{mean}(D^v)$	0.8	0.684	0.800	0.973	0.828

D. Discussion

1) Breast Density: We investigated the patient density model D and aggregated view model $\text{mean}(D^v)$ in depth and varied the decision threshold (see Table VIII). The results show that model D yields more reliable predictions with high confidence, and thus, higher sensitivities, accuracies, and F1 scores at various thresholds compared to the aggregated view model ($p < 0.001$). Such automated tools that deliver trustworthy, reproducible measures are of increasing importance in clinical practice, especially for breast density assessment where subjectivity and high inter-observer variability are well-known issues [2], [6], [23]. As breast density is considered an important risk factor for the development of breast cancer, reliability and reproducibility are key aspects when it comes to standardized density reporting which may trigger supplemental/personalized screening procedures [2], [23], [69].

2) Comparison to Related Work: Table IX sets our method in context to related approaches in the literature. In general, a direct comparison of reported evaluation measures of different methods is not possible as datasets used for training and evaluation differ vastly, e.g., varying imaging quality and modality (scanned film vs. full-field digital mammography), overall number of images, or amount of training data. To counteract this issue at least to some extent, we report train and test databases in Table IX and refer to the respective publications for further details. In addition, we compare our results to those reported with a single (fusion) model and without augmentation at test time.

Fusion-based methods: Overall, our multi-input CNNs improve AUC scores by 0.029 to 0.04 compared to naive model ensembling. Similar increases for fusion approaches have also been reported in the literature. Kooi et al. [4] report improved AUC (+ 0.019) when adding handcrafted features (like contrast, texture) to CNN features for the classification of single mammograms. The work by Kyono et al. [52] fuses multi-task scores, like "diagnosis", "suspicion", "conspicuity",

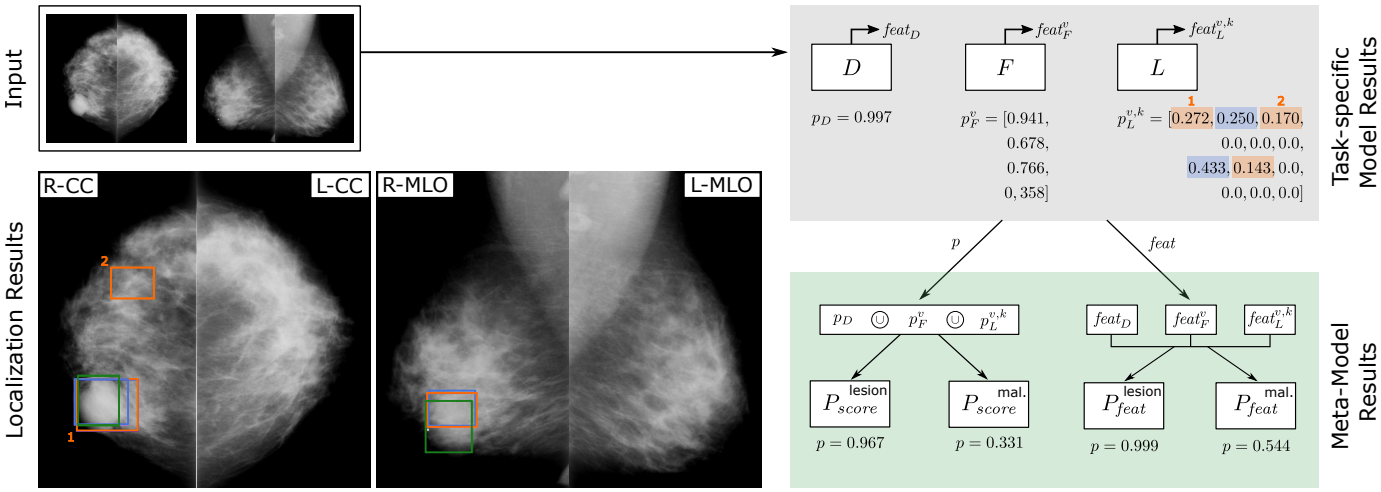


Fig. 12: Illustration of the benefits of our mammography pipeline: The patient has a malignant mass (green = ground truth) in the right breast. Model L was able to localize the malignant mass (orange), but with low confidence that is not sufficient to be reliably counted as detection. Low confidence localizations were also found by L for an additional malignant mass and a benign mass (blue). Detection scores $p_L^{v,k}$ are colored according to their class label as detected by L. Results show that both fusion models P_{feat} are able to circumvent the low scores and correctly classify the patient (malignancy = 0.544, lesion = 0.999). Best viewed in color.

TABLE IX: Results obtained with fusion models P_{feat} compared to classification results reported in related works. Train and test datasets utilized by the respective methods are separated with "/".

Method	Train/Test Data	Fusion-Level	Result-Level	Target	AUC	AUPRC
Shen [43]	CBIS-DDSM / CBIS-DDSM	-	image	mal.	0.87	-
	CBIS-DDSM + INbreast / INbreast	-	image	mal.	0.95	-
Shu [47]	CBIS-DDSM / CBIS-DDSM	-	image	mal.	0.838	-
	INbreast / INbreast	-	image	mal.	0.934	-
Ribli [3]	DDSM, private / INbreast	-	breast (max+avg)	mal.	0.95	-
Lotter [42]	DDSM, OPTIMAM, private / OPTIMAM	ROI	patient (avg+max)	mal.	0.963 $\pm$ 0.003	-
Kooi [4]	private / private (NL screening)	ROI	image	mal. mass	0.941	-
Shen [18]	CBIS-DDSM / CBIS-DDSM	image	breast (avg)	mal.	0.833	-
	private / private (NYU)	image	breast (avg)	mal.	0.891	0.390
Shachor [37]	DDSM / DDSM	ROIs (from CC+MLO)	breast	ben./mal. calc.	0.661	-
Kyono [52]	private / private (Tommy trial)	patient	patient	mal.	0.824 $\pm$ 0.016	0.580 $\pm$ 0.028
Ours (P_{feat})	DDSM / DDSM	patient	patient	mal.	0.791	0.660
		patient	patient	lesion	0.962	0.987

"breast density", across multiple views, similar to our method, but with the difference that their multi-task model predicts the same scores per view image, while we fuse predictions obtained from *different* models. Adding the multi-task output to their multi-view approach increased performance by 0.031 in terms of AUC. Shen et al. [18] fuse information on a single-image level in a weakly-supervised fashion, i.e., by fusing salient image regions with a fusion module, and report a single-model AUC score of 0.833 on CBIS-DDSM test data. A recent method by Lotter et al. [42] combines fully and weakly (multi-instance) supervised learning and claims state-of-the-art performance for mammogram classification (AUC = 0.963 $\pm$ 0.003, OPTIMAM data). However, to obtain a score on patient level, they perform standard ensembling (average + maximum). Finally, McKinney et al. [7] average cancer risk scores that are predicted by an ensemble of three large-scale deep learning models (AUC = 0.889, OPTIMAM data). Each model fuses features at different stages and aggregates predictions in various ways, e.g., by considering the maximum score or via MLPs.

When looking at the dedicated lesion prediction, we observe that - to the best of our knowledge - our method is the only one that specifically investigated this classification target. With an AUC score of 0.962 (F1 = 0.948), this model could be reliably used, e.g., within a reporting system, where patients with lesions are examined first.

Non-fusion-based methods: Apart from the summarized information fusion methods, there are numerous works that predict whether an image is malignant directly from a view image [43], [47], or/and additionally apply simple ensembling strategies for predictions on breast- or patient-level [3], [13]. Shen et al. [43], for example, utilize patch-based pre-training and compare variants of ResNet and VGG in their work. They report an image-level AUC score of 0.87 on CBIS-DDSM and 0.95 on the INbreast dataset (transfer-learned). Shu et al. [47] propose two region-based pooling strategies

and achieve lower AUC scores on CBIS-DDSM (0.838) and INbreast (0.934) data as compared to Shen et al. [43]. Ribli et al. [3] localize suspicious lesions using Faster R-CNN and consider the maximum/average score on image/breast level (AUC = 0.95 on INbreast data).

Fusion- vs. non-fusion-based methods: The results summarized in Table IX show competitive performance of fusion- and non-fusion-based methods. Although there is no clear benefit of fusion-based approaches over non-fusion-based works in terms of AUC scores, fusion approaches show different advantages. Recent methods focused, for example, on the integration of radiological and clinical features or aimed at increasing interpretability of models, which is an important aspect in the medical domain [11], [18], [42], [52]. These advantages, however, may come at the cost of more complex training procedures as compared to standard deep learning models [18]. One limitation of recent fusion-based approaches, including this work, is the requirement for detailed, high-quality expert-annotations [4], [7], [9], [11]. However, this is not limited to fusion methods per se, as the need for, e.g., bounding box annotations applies likewise to non-fusion-based, R-CNN/YOLO-based localization approaches [3], [29], [39], [41]. Recent weakly supervised works aim to tackle this issue and show already promising results [18], [42], [48].

3) Clinical Implications: In this work, we presented a technical proof-of-concept study for a mammography pipeline comprising of three task-specific models and patient meta-models that fuse task-specific features and predictions. While one goal was to obtain an improved assessment on patient level as compared to standard model ensembling, the second goal was to develop a support tool for reading tasks of radiologists. Similar to recent technical proof-of-concept studies by Kyono et al. [9], [52] and Barnett et al. [11], we aimed to provide intermediate results that are linked to radiological features and potential cancer risk factors. This is in contrast to studies that highlight the potential of workload reduction by excluding scans from reading that are very likely normal, i.e., do not have any suspicious lesions [9], [10], [15], [70]. However, our global lesion and malignancy predictions could be used to prioritize images for reading instead of excluding them, and additional intermediate results of task-specific models can be presented to the clinicians during exam reading and diagnosis. Localizing suspicious lesions, for example, is an essential part when reading mammograms where clinicians examine both views and breasts [37]. Thus, showing localized regions can aid radiologists in image interpretation [71], for example, showing only the most important findings and raising attention for them [14], [72]. Recent user studies in other domains like prostate cancer diagnosis confirmed that prompting clinicians to suspicious regions helped them in reading [20]. In addition to the localized regions, our pipeline estimates the patient's breast density, which is an important risk factor for developing breast cancer [2], as already discussed in Section IV-D.1.

4) Limitations: One limitation of this study is the relatively small DDSM dataset with only 2254 patients (after curation), as compared to resources used by related works [7], [18], [24], [41], [52]. Further, as the DDSM data consists of scanned film mammograms only, the imaging quality is significantly lower as compared to full-field digital mammography images. However, the usage of a fully open dataset fosters the development and comparability of approaches, while, e.g., access to the OPTIMAM database and high-quality, expert-annotated data in general remains limited [10], [71]. A second factor may be the fixed number of detected lesions n currently used in our fusion models, which could be targeted with a multi-instance approach in the future. Finally, the use of a combined model that performs lesion and malignancy prediction in a multi-task fashion would reduce the number of models and could potentially improve also the performance of malignancy prediction.

5) Future Perspectives: The transfer of the complete pipeline to a large, full-field digital mammography dataset would be the logical next step and potentially boost performance when trained on a larger data resource. To mitigate the requirement for expensive bounding box annotations for the lesion localization model L, an interpretable, weak localization approach similar to our recent works [59], [60] can be integrated in conjunction with the findings model F. Further, the improvement of the localization performance of model L, especially for benign lesions, as well as the training of one combined model for lesion and malignancy prediction are considered future work. The inclusion of additional radiological features, such as non-image-based risk factors (e.g., patient age, patient/family history) would be of interest and importance for clinical use [73]. Moreover, the analysis of temporal change of lesions is considered an important biomarker in practice [21], [27]. Finally, the evaluation of our proposed pipeline in a clinical reader study would help to further evaluate the potential benefits and risks of having global and local, task-specific information available, e.g., in terms of acceptance, increased interpretability, but also potential bias [10].

V. Conclusion

In this work, we proposed the fusion of predictions and features from different *task-specific models* for improving mammography screening data classification. We trained and evaluated our fusion models for two different classification targets relevant in the field of mammogram analysis: the prediction of *(i)* the presence of any lesion and *(ii)* the presence of any malignant lesion in a patient. Our experiments on public mammography data showed that the fusion of scores with MLPs as well as feature fusion with multi-input embedding CNNs improves AUC scores compared to standard ensembling. Overall, we report an AUC score of 0.962 for predicting the presence of any lesion and 0.791 for classifying the presence of malignant lesions on patient level. By supporting our global predictions per patient with the local, sub-results obtained by the task-specific models, we aim to aid clinicians in their reading and decision process. Finally, we performed an ablation study with breast density scores and features and conclude that additional density information can benefit the classification performance for both target scores.

REFERENCES

[1] C. P. Wild, E. Weiderpass, and B. W. Stewart, *World Cancer Report 2020: Cancer research for cancer precention*. Lyon: International Agency for Research on Cancer, 2020.

[2] S. V. Destounis, A. Santacroce, and A. Arieno, "Update on breast density, risk estimation, and supplemental screening," *AJR Am. J. Roentgenol.*, vol. 214, no. 2, pp. 296–305, 2020.

[3] D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, "Detecting and classifying lesions in mammograms with deep learning," *Sci. Rep.*, vol. 8, no. 1, p. 4165, 2018.

[4] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer, "Large scale deep learning for computer aided detection of mammographic lesions," *Med. Image Anal.*, vol. 35, pp. 303–312, 2017.

[5] C. D. Lehman, A. Yala, T. Schuster, B. Dontchos, M. Bahl, K. Swanson, and R. Barzilay, "Mammographic breast density assessment using deep learning: Clinical implementation," *Radiology*, vol. 290, no. 1, pp. 52–58, 2019.

[6] N. Kaiser, A. Fieselmann, S. Vesal, N. Ravikumar, L. Ritschl, S. Kappler, and A. Maier, "Mammographic breast density classification using a deep neural network: Assessment based on inter-observer variability," in *Proc. SPIE Med. Imag.*, vol. 10952. SPIE, 2019, pp. 156 – 161.

[7] S. M. McKinney *et al.*, "International evaluation of an AI system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[8] A. Yala, C. Lehman, T. Schuster, T. Portnoi, and R. Barzilay, "A deep learning mammography-based model for improved breast cancer risk prediction," *Radiology*, vol. 292, no. 1, pp. 60–66, 2019.

[9] T. Kyono, F. J. Gilbert, and M. van der Schaar, "Improving workflow efficiency for mammography using machine learning," *J. Am. Coll. Radiol.*, vol. 17, pp. 56–63, 2020.

[10] W. C. Ou, D. Polat, and B. E. Dogan, "Deep learning in breast radiology: current progress and future directions," *Eur. Radiol.*, vol. 31, no. 7, pp. 4872–4885, 2021.

[11] A. J. Barnett, F. R. Schwartz, C. Tao, C. Chen, Y. Ren, J. Lo Y, and C. Rudin, "Interpretable mammographic image classification using cased-based reasoning and deep learning," in *Proc. IJCAI Workshop on Deep Learning, Case-Based Reasoning, and AutoML: Present and Future Synergies.*, 2021. [Online]. Available: http://arxiv.org/pdf/2107.05605v1

[12] T. Schaffter *et al.*, "Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms," *JAMA Netw. Open*, vol. 3, no. 3, p. e200265, 2020.

[13] H.-E. Kim *et al.*, "Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study," *Lancet Digit. Health*, vol. 2, no. 3, pp. 138–148, 2020.

[14] A. Rodríguez-Ruiz, E. Krupinski, J.-J. Mordang, K. Schilling, S. H. Heywang-Köbrunner, I. Sechopoulos, and R. M. Mann, "Detection of breast cancer with mammography: Effect of an artificial intelligence support system," *Radiology*, vol. 290, no. 2, pp. 305–314, 2019.

[15] K. J. Geras, R. M. Mann, and L. Moy, "Artificial intelligence for mammography and digital breast tomosynthesis: Current concepts and future perspectives," *Radiology*, vol. 293, no. 2, pp. 246–259, 2019.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1135–1144.

[17] A. Barredo Arrieta *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inform. Fusion*, vol. 58, pp. 82–115, 2020.

[18] Y. Shen *et al.*, "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization," *Med. Image Anal.*, vol. 68, p. 10908, 2021.

[19] P. Tschandl *et al.*, "Human-computer collaboration for skin cancer recognition," *Nat. Med.*, vol. 26, no. 8, pp. 1229–1234, 2020.

[20] C. J. Cai, S. Winter, D. Steiner, L. Wilcox, and M. Terry, "'Hello AI': Uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, pp. 1–24, 2019.

[21] I. Sechopoulos, J. Teuwen, and R. Mann, "Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art," *Semin. Cancer Biol.*, vol. 72, pp. 214–225, 2021.

[22] E. A. Sickles *et al.*, "ACR BI-RADS® Mammography," in *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System*. Reston, VA: American College of Radiology, 2013.

[23] B. L. Sprague *et al.*, "Variation in mammographic breast density assessments among radiologists in clinical practice: Findings from a multicenter observational study," *Ann. Intern. Med.*, vol. 165, no. 7, pp. 457–464, 2016.

[24] N. Wu *et al.*, "Breast density classification with deep convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 6682–6686.

[25] A. A. Mohamed, W. A. Berg, H. Peng, Y. Luo, R. C. Jankowitz, and S. Wu, "A deep learning method for classifying mammographic breast density categories," *Med. Phys.*, vol. 45, no. 1, pp. 314–321, 2018.

[26] M. Kallenberg *et al.*, "Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1322–1331, 2016.

[27] T. Kooi and N. Karssemeijer, "Classifying symmetrical differences and temporal change for the detection of malignant masses in mammography using deep neural networks," *J. Med. Imaging*, vol. 4, no. 4, p. 044501, 2017.

[28] G. Carneiro, J. Nascimento, and A. P. Bradley, "Automated analysis of unregistered multi-view mammograms with deep learning," *IEEE Trans. Med. Imag.*, vol. 36, no. 11, pp. 2355–2365, 2017.

[29] A. Akselrod-Ballin, L. Karlinsky, A. Hazan, R. Bakalo, A. B. Horesh, Y. Shoshan, and E. Barkan, "Deep learning for automatic detection of abnormal findings in breast mammography," in *Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, ser. Lecture Notes in Computer Science, vol. 10553. Springer Int. Publishing, 2017, pp. 321–329.

[30] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. G. Lopez, "Representation learning for mammography mass lesion classification with convolutional neural networks," *Comput. Methods Programs Biomed.*, vol. 127, pp. 248–257, 2016.

[31] R. K. Samala, H.-P. Chan, L. M. Hadjiiski, M. A. Helvie, K. H. Cha, and C. D. Richter, "Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms," *Phys. Med. Biol.*, vol. 62, no. 23, pp. 8894–8908, 2017.

[32] J.-J. Mordang, T. Janssen, A. Bria, T. Kooi, A. Gubern-Mérida, and N. Karssemeijer, "Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks," in *Breast Imaging*, ser. Lecture Notes in Computer Science, vol. 9699. Springer Int. Publishing, 2016, pp. 35–42.

[33] Y. Dgani, H. Greenspan, and J. Goldberger, "Training a neural network based on unreliable human annotation of medical images," in *Proc. IEEE Int. Symp. Biomed. Imag. (ISBI)*. IEEE, 2018, pp. 39–42.

[34] I. El-Naqa, Y. Yang, M. N. Wernick, N. P. Galatsanos, and R. M. Nishikawa, "A support vector machine approach for detection of microcalcifications," *IEEE Trans. Med. Imag.*, vol. 21, no. 12, pp. 1552–1563, 2002.

[35] R. Agarwal, O. Diaz, X. Lladó, M. H. Yap, and R. Martí, "Automatic mass detection in mammograms using deep convolutional neural networks," *J. Med. Imaging*, vol. 6, no. 3, p. 031409, 2019.

[36] J. Wang and Y. Yang, "A context-sensitive deep learning approach for microcalcification detection in mammograms," *Pattern Recognit.*, vol. 78, pp. 12–22, 2018.

[37] Y. Shachor, H. Greenspan, and J. Goldberger, "A mixture of views network with applications to multi-view medical imaging," *Neurocomputing*, vol. 374, pp. 1–9, 2020.

[38] N. Dhungel, G. Carneiro, and A. P. Bradley, "A deep learning approach for the analysis of masses in mammograms with minimal user intervention," *Med. Image Anal.*, vol. 37, pp. 114–128, 2017.

[39] M. A. Al-Masni *et al.*, "Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system," *Comput. Methods Programs Biomed.*, vol. 157, pp. 85–94, 2018.

[40] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[41] R. Agarwal, O. Díaz, M. H. Yap, X. Lladó, and R. Martí, "Deep learning for mass detection in full field digital mammograms," *Comput. Biol. Med.*, vol. 121, p. 103774, 2020.

[42] W. Lotter *et al.*, "Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach," *Nat. Med.*, vol. 27, no. 2, pp. 244–249, 2021.

[43] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," *Sci. Rep.*, vol. 9, no. 1, p. 12495, 2019.

[44] W. Lotter, G. Sorensen, and D. Cox, "A multi-scale CNN and curriculum learning strategy for mammogram classification," in *Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, ser. Lecture Notes

in Computer Science, vol. 10553. Springer Int. Publishing, 2017, pp. 169–177.

[45] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, vol. 10435. Springer Int. Publishing, 2017, pp. 603–611.

[46] N. Wu *et al.*, "Deep neural networks improve radiologists' performance in breast cancer screening," *IEEE Trans. Med. Imag.*, vol. 39, no. 4, pp. 1184–1194, 2020.

[47] X. Shu, L. Zhang, Z. Wang, Q. Lv, and Z. Yi, "Deep neural networks with region-based pooling structures for mammographic image classification," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 2246–2255, 2020.

[48] M. Tardy and D. Mateus, "Looking for abnormalities in mammograms with self-and weakly supervised reconstruction," *IEEE Trans. Med. Imag.*, 2021.

[49] Y. Shen, N. Wu, J. Phang, J. Park, G. Kim, L. Moy, K. Cho, and K. J. Geras, "Globally-aware multiple instance classifier for breast cancer screening," in *Proc. Mach. Learn. Med. Imag.*, ser. Lecture Notes in Computer Science, vol. 11861, 2019, pp. 18–26.

[50] A. Jouirou, A. Baâzaoui, and W. Barhoumi, "Multi-view information fusion in mammograms: A comprehensive overview," *Inform. Fusion*, vol. 52, pp. 308–321, 2019.

[51] K. J. Geras *et al.*, "High-resolution breast cancer screeining with multi-view deep convolutional neural networks," arXiv:1703.07047, 2018.

[52] T. Kyono, F. J. Gilbert, and M. van der Schaar, "Multi-view multi-task learning for improving autonomous mammogram diagnosis," in *Proc. Mach. Learn. Healthcare*, ser. Proceedings of Machine Learning Research, vol. 106, 2019, pp. 571–591.

[53] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: efficient convolutional neural networks for mobile vision applications," arxiv:1704.04861, 2017.

[54] M. Heath, K. Bowyer, D. Kopans, P. Kegelmeyer, JR, R. Moore, K. Chang, and S. Munishkumaran, "Current status of the digital database for screening mammography," in *Digit. Mammography*, ser. Computational Imaging and Vision, vol. 13, 1998, pp. 457–460.

[55] M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, "The digital database for screening mammography," in *Proc. Int. Workshop Digit. Mammography*. Medical Physics Publishing, 2001, pp. 212–218.

[56] R. S. Lee, F. Gimenez, A. Hoogi, and D. Rubin, "Curated breast imaging subset of DDSM [dataset]," The Cancer Imaging Archive.

[57] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Sci. Data*, vol. 4, p. 170177, 2017.

[58] K. Clark *et al.*, "The cancer imaging archive (TCIA): maintaining and operating a public information repository," *J. Digit. Imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[59] D. Major, D. Lenis, M. Wimmer, G. Sluiter, A. Berg, and K. Bühler, "Interpreting medical image classifiers by optimization based counterfactual imapct analysis," in *Proc. IEEE Int. Symp. Biomed. Imag. (ISBI)*. IEEE, 2020, pp. 1096–1100.

[60] D. Lenis, D. Major, M. Wimmer, A. Berg, G. Sluiter, and K. Bühler, "Domain aware medical image classifier interpretation by counterfactual impact analysis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, ser. Lecture Notes in Computer Science, vol. 12261. Springer, 2020, pp. 315–325.

[61] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2818–2826.

[62] F. Chollet *et al.*, "Keras," 2015. [Online]. Available: https://keras.io

[63] M. Abadi *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: https://www.tensorflow.org/

[64] J. Huang *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 7310–7311.

[65] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," in *Proc. Uncertain. Artif. Intell.*, 2018, pp. 876–885.

[66] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017.

[67] A. Oliver, J. Freixenet, R. Martí, J. Pont, E. Pérez, E. R. E. Denton, and R. Zwiggelaar, "A novel breast tissue density classification methodology," *IEEE Trans. Inf. Technol. Biomed.*, vol. 12, no. 1, pp. 55–65, 2008.

[68] J. Anitha, J. D. Peter, and S. I. A. Pandian, "A dual stage adaptive thresholding (DuSAT) for automatic mass detection in mammograms," *Comput. Methods Programs Biomed.*, vol. 138, pp. 93–104, 2017.

[69] E. F. Conant, B. L. Sprague, and D. Kontos, "Beyond BI-RADS density: A call for quantification in the breast imaging clinic," *Radiology*, vol. 286, no. 2, pp. 401–404, 2018.

[70] A. Rodriguez-Ruiz *et al.*, "Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? a feasibility study," *Eur. Radiol.*, vol. 29, no. 9, pp. 4825–4832, 2019.

[71] S. Soffer, A. Ben-Cohen, O. Shimon, M. M. Amitai, H. Greenspan, and E. Klang, "Convolutional neural networks for radiologic images: A radiologist's guide," *Radiology*, vol. 290, no. 3, pp. 590–606, 2019.

[72] A. Rodriguez-Ruiz *et al.*, "Stand-alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists," *J. Natl. Cancer Inst.*, vol. 111, no. 9, pp. 916–922, 2019.

[73] O. Weaver and J. W. T. Leung, "Biomarkers and imaging of breast cancer," *AJR Am. J. Roentgenol.*, vol. 210, no. 2, pp. 271–278, 2018.