MULTI-VIEW DEEP EVIDENTIAL FUSION NEURAL NETWORK FOR ASSESSMENT OF SCREENING MAMMOGRAMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mammography is an X-ray-based imaging technique widely used for breast cancer screening and early-risk assessment. A large number of mammograms are acquired in regular breast cancer screening programs. The assessment of mammograms is a tedious task and may be difficult to accomplish due to a shortage of expert radiologists in breast imaging. Artificial intelligence-powered algorithms, especially deep learning, could assist radiologists by automating the assessment, however, substantial trust needs to be established in incorporating such algorithms in real-world settings. The evidential neural networks algorithm provides an interpretable approach using Dempster-Shafter evidential theory that supports network predictive confidence. Recent studies have suggested that multi-view analysis improves the assessment of mammograms. In this study, we advance the multi-view assessment of mammograms by using a deep evidential neural network to address the following questions:

1. What is the effect of various pre-trained convolutional neural networks in extracting features from mammograms?
2. Which fusion strategies work better for the multi-view assessment of mammograms using a deep evidential learning framework?

The multi-view deep evidential neural network extracts features from each mammogram's view using a pre-trained convolutional neural network. The extracted features are combined using Dempster-Shafer evidence theory for the following two classification tasks, mammogram density assessment in BI-RADS categories and mammogram finding as benign or malignant. We conducted extensive experiments using two open-sourced digital mammogram datasets, VinDr-mammo, and mini-DDSM, with 4,977 and 1,885 patients, each with four mammogram views, respectively. The results suggest that the multi-view approach outperforms the single-view by relative improvements of 2.99% and 2.64% for VinDr-mammo, and 6.51% and 8.75% for mini-DDSM datasets, in terms of F1-score, in mammogram density assessment and BI-RADS findings benign/malignant classification tasks, respectively. Our results show that the multi-view assessment of mammograms using a deep evidential fusion approach not only provides superior performance than the single-view assessment but also enhances trust in incorporating artificial intelligence-powered algorithms for the assessment of screening mammograms.

1 INTRODUCTION

Breast cancer (BC) is the women's most common and leading cancer type, with an estimated 290,560new cases in United States, 2022 (Siegel et al., 2022). Early detection of BC is critical to lower the BC mortality rate and healthcare costs. Breast screening techniques, especially mammography, are utilized for early BC detection. A typical mammogram consists of four views: left craniocaudal (L-CC), right craniocaudal (R-CC), left mediolateral oblique (L-MLO), and right mediolateral oblique (R-MLO). In a standard screening process, a radiologist qualitatively assesses mammograms and reports the findings according to the Breast Imaging Reporting and Data System (BI-RADS) (D'Orsi et al., 2013). BI-RADS scoring is widely accepted to report mammogram density

assessment and mammogram findings. The BI-RADS score contains seven assessment categories: BI-RADS 0 (incomplete), BI-RADS 1 (negative), BI-RADS 2 (benign), BI-RADS 3 (probably benign), BI-RADS 4 (suspicious for malignancy), BI-RADS 5 (highly suggestive of malignancy), and BI-RADS 6 (known biopsy-proven malignancy). In addition to the BI-RADS score, radiologists report the breast density, which is the relative amount of fibroglandular tissues within the breast area, by examining the mammograms. The breast density assessment contains four categories, including A (almost entirely fatty breast), B (scattered areas of fibroglandular density), C (heterogeneously dense breast), and D (extremely dense) (D'Orsi et al., 2013). The European Commission Initiative on Breast Cancer (ECIBC) strongly recommends organizing regular mammography screening programs for the early detection of breast cancer in asymptomatic women. The assessment of the enormous number of screening mammograms is challenging to accomplish due to the shortage of expert breast imaging radiologists. Artificial intelligence (AI) approaches, specifically deep learning (DL) methods, could assist radiologists by automating the assessment of screening mammograms.

Deep learning algorithms have been widely used for tasks, such as mass detection (Dhungel et al., 2015), micro-classification (Wang et al., 2016), and breast density assessment (Shen et al., 2019) using mammograms. Recent studies have suggested that multi-view-based approaches improve the performance of such tasks (Wu et al., 2019; Khan et al., 2019; Geras et al., 2017; Li et al., 2020; Seyyedi et al., 2020; Nguyen et al., 2022b). In general, multi-view approaches extract the imaging features from all four mammogram views and combine them for various downstream tasks. Feature extraction is the most crucial step; the behavior of the predictive models depends on the extracted features. Pre-trained convolutional neural networks (CNNs), trained on Imagenet (Deng et al., 2009), have been applied successfully to various image recognition tasks as feature extractors or as a backbone architecture for transfer learning (Penatti et al., 2015; Azizpour et al., 2015; Tajbakhsh et al., 2016). However, a best-performing pre-trained model should be chosen from many available models for each downstream task. Moreover in mammogram-based classification tasks, the gap still exists in selecting the best multi-view fusion strategies, i.e., fusion at feature or decision levels.

Incorporating multi-view fusion strategies for mammogram assessment in a radiology practice requires trust. Developing trust-aware and robust models is crucial in the medical domain. Emerging evidential deep learning (EDL) explores cutting-edge possibilities to develop trust-aware AI models. Dempster-Shafter theory of evidence (DST) (Shafer, 1976) mathematically provides an end-to-end interpretable approach to quantify the uncertainty in the model predictions. The Dirichlet distribution generates the evidence from the extracted features, then the subjective logic (Jøsang, 2016) formalizes the evidence to belief masses that support the model decision. Inspired by the EDL (Sensoy et al., 2018) and information fusion (Tong et al., 2021; Han et al., 2021) strategies, we develop a trust-aware and robust model for the assessment of screening mammograms in the following two image classification tasks, mammogram density assessment in terms of BI-RADS categories and mammogram's finding as benign or malignant classes. We conduct extensive experiments to investigate the effect of using various pre-trained models to extract mammogram features and compare different evidential fusion strategies.

2 METHODOLOGY

2.1 PRE-TRAINED CONVOLUTIONAL NEURAL NETWORKS

Convolutional neural networks (CNNs) have been incorporated for various tasks in the medical domain. The depth of the network plays a vital role in extracting high-, mid-, and low-level features from the given input image. Increase in the depth of the CNNs poses a challenging to train the models (He et al., 2016). Recently, CNNs pre-trained on ImageNet weights (Deng et al., 2009) have been widely utilized to extract the input features. Transfer learning and fine-tuning techniques work well in the medical domain (Raghu et al., 2019); however, there is inconsistency in adopting pre-trained CNN models for mammogram classification task (Wang et al., 2020). In this study, we investigate different pre-trained CNN models for two downstream mammogram classification tasks, namely mammogram density assessment and mammogram's BI-RADS findings as benign or malignant. In this study, we selected five pre-trained CNN architectures, including VGG (Simonyan & Zisserman, 2014), Resnet (He et al., 2016), Densenet (Huang et al., 2017), Inceptionnet (Szegedy

et al., 2016), and Efficientnet (Tan & Le, 2019), to extract mammogram features. We investigated 29 pre-trained CNN variants from these five CNN architectures.

2.2 DEMPESTER-SHAFTER EVIDENCE THEORY

Convolutional neural networks generally use the SoftMax activation function (Szandała, 2021) at the output layer to estimate the class probabilities. However, the SoftMax predictions lead to overconfidence (Moon et al., 2020). To quantify the uncertainty in the output predictions, different techniques, including Monte-Carlo dropout regularization (Gal & Ghahramani, 2016), deep ensemble networks (Rahaman et al., 2021), and test-time augmentations (Wang et al., 2019), have been proposed. These techniques require additional sampling processes and are computationally expensive (Sensoy et al., 2018). In contrast to the existing approaches, DST provides an interpretable approach to quantify the evidence and the total predictive uncertainty within the same neural network model. The extracted features at the CNN output layer are sampled using Dirichlet distribution with concentration parameter α. The subjective logic then formalizes the distribution as belief masses based on the evidence of each classification category and an overall uncertainty associated with the network. For the classification task with K mutually exclusive classes, the acquired subjective belief masses are all non-negative, and their sum is equal to 1, defining as below (Han et al., 2021):

$$u + \sum_{k}^{K} b_k = 1 \tag{1}$$

where, $u \geq 0$ and $b_k \geq 0$ denote the overall uncertainty and belief mass of the k^{th} class, respectively.

The acquired associated evidence e_k from the input to the classification support, is closely related to the expected concentration parameters α_k , specifically related as $e_k = \alpha_k - 1$. Accordingly, the belief masses are computed as (Han et al., 2021):

$$b_k = \frac{e_k}{S}, \quad u = \frac{K}{S} \tag{2}$$

Where $S = \sum_{k=1}^{K}(e_k + 1) = \sum_{k=1}^{K} \alpha_k$ is known as the Dirichlet strength. From eq. 2, the more evidence obtained for the k^{th} class, the higher the assigned belief mass will be.

Mathematical intuition behind the EDL and the end-to-end trainable loss function details are provided in appendix A.

2.3 EVIDENTIAL FUSION STRATEGIES

Recent studies suggested that combining multiple mammogram views enhances the model prediction accuracy (Khan et al., 2019). In general, the fusion strategies include two categories (Ilhan et al., 2022); feature-level fusion, where the mammogram features from each view are concatenated before the classification and late fusion, where the mammogram-view classification network decision are combined. In this study, we investigated feature-level and late-fusion strategies at the mammogram and view-specific levels. Figure 1 illustrates various fusion strategies considered in this study. In the feature-level fusion strategies, the extracted mammogram features from the pre-trained CNN model are concatenated and then the combined features pass through the evidential layer. In late-fusion strategy, the extracted mammogram features are first passed through the evidential layers and then belief mass of each mammogram are combined using Dempster's combination rule (Han et al., 2021):

The investigated evidential fusion strategies are:

1. Feature-level evidential fusion (FLEF): The mammogram features of all the four-view are extracted independently using a pre-trained CNN and then concatenated. The concatenated features are passed through the evidential layer consisting of variational Dirichlet distribution and subjective belief mass of each category.
2. View-specific feature-level evidential fusion (VS-FLEF): The mammogram features of the right and left side are concatenated, and then the combined view-specific features are passed through the evidential layers. The view-specific subjective belief masses are combined using the Dempster's combination rule.

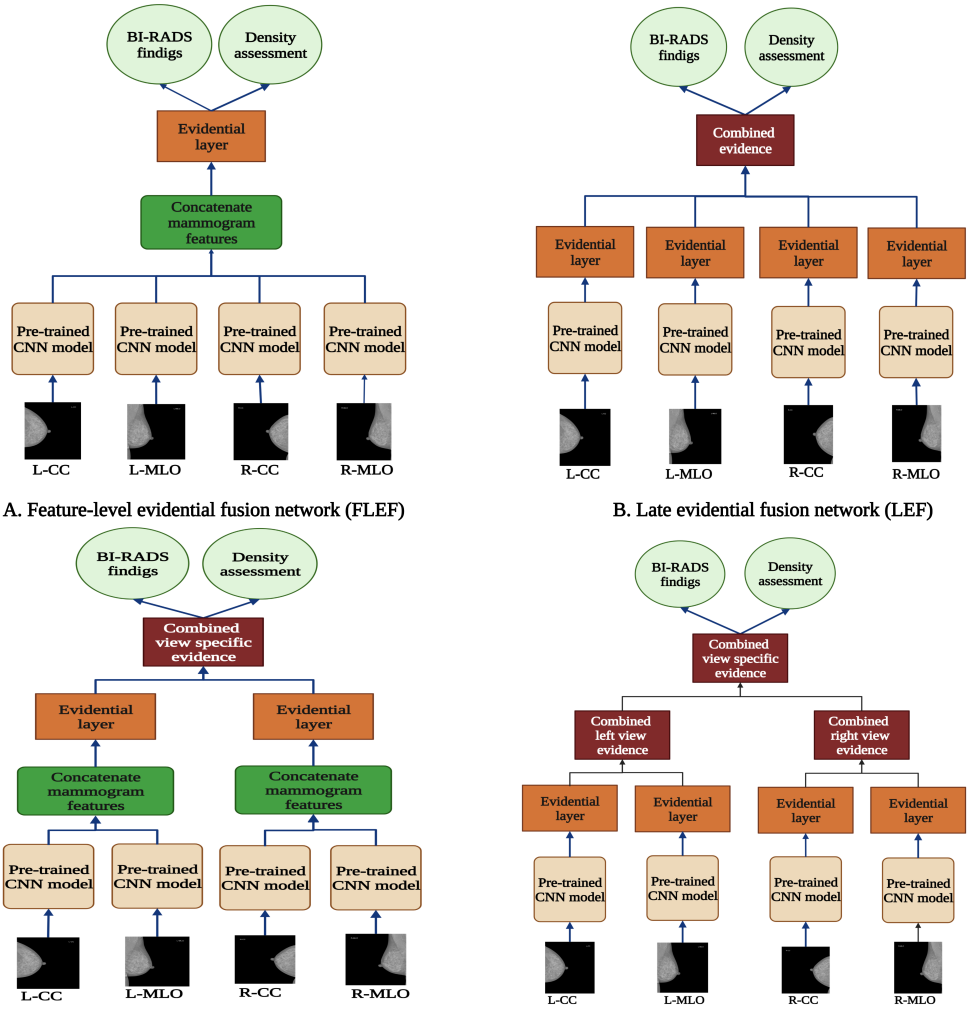

Figure 1: Different evidential fusion strategies considered in this study. The sub-figures A and B represent mammogram feature-level, and late evidential fusion strategies, and the sub-figures C and D represent the view-specific mammogram feature-level and view-specific late evidential fusion strategies. The features extracted from the pre-trained CNN models are concatenated in the feature-level fusion strategies, followed by evidential learning. In late fusion strategies, the extracted mammogram features are passed through the evidential layer, and then the evidences are combined using Dempster's combination rule.

3. Late evidential fusion (LEF): The extracted mammogram features of each view are independently passed through the evidential layers. Then the subjective belief masses of each view are combined using the Dempster's combination rule.

4. View-specific late evidential fusion (VS-LEF): The extracted mammogram features of each view are independently passed through the evidential layers. The subjective belief masses of each side are combined, followed by combining the view-specific subjective belief masses using the Dempster's combination rule.

3 DATASETS

We used two open-sourced digital mammogram datasets, namely VinDr-mammo (Nguyen et al., 2022a) and mini-DDSM (Lekamlage et al., 2020), with 5,000 and 1,975 patients, respectively. The datasets are provided with BI-RADS scoring and density annotations. For simplicity, based on the supervision of an expert radiologist, we categorized BI-RADS density scores into two separate categories: BI-RADS 2 and 3 as benign and BI-RADS 5 and 6 as malignant. Table 1 shows the distribution of patients in VinDr-mammo and mini-DDSM datasets for mammogram density assessment in terms of BI-RADS categories and mammogram findings as benign or malignant classes. Note that we only considered the patient's data having all four views of the mammogram. For the mammogram BI-RADS findings task, the number of patients is less than the density assessment task, as we considered only the benign and malignant cases.

Table 1: Distribution of mammograms in the VinDr-mammo and mini-DDSM datasets for the density assessment and BI-RADS findings tasks.

Dataset	**Mammogram density assessment**			
	Density A	**Density B**	**Density C**	**Density D**
VinDr-mammo (4977)	24	477	3807	669
mini-DDSM (1885)	264	710	568	343
Dataset	**Mammogram BI-RADS findings**			
	Benign		**Malignant**	
VinDr-mammo (4263)	3178		1085	
mini-DDSM(1350)	671		679	

3.1 DATA PRE-PROCESSING

Most mammograms have a black background with view labels. We pre-processed the mammograms using a minimum bounding box which captures only the mammogram and then, normalized each mammogram using a min-max scaler. The presence of pectoral muscle in the MLO-view mammograms will disrupt mammogram diagnosis and may also result in a false positive diagnosis (Cardoso et al., 2010). Thus, detecting the pectoral muscle and its removal is essential for mammogram diagnosis. Therefore, we employed a breast segmentation algorithm proposed by Gudhe et al. (2022) to segment the pectoral muscle from mammograms. The pre-trained CNN models used in this study have different image resolutions. We employed the default image resolution of each pre-trained CNN architecture and resized it using the bicubic interpolation technique (Han, 2013). Figure 2 illustrates the pre-processing steps for a randomly-selected MLO-view mammogram from the mini-DDSM dataset.

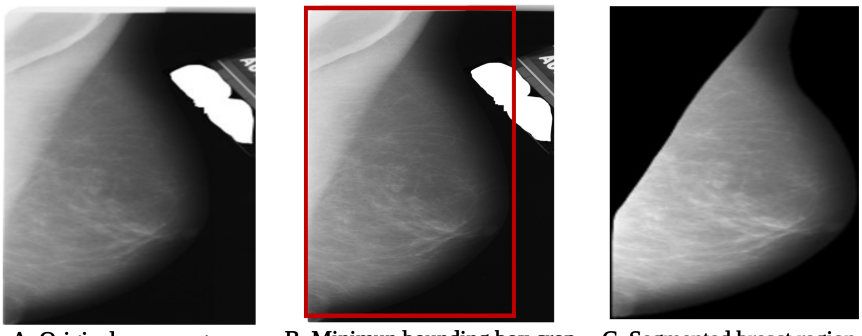

Figure 2: Pre-processing steps for a randomly selected mammogram from the mini-DDSM dataset. Sub-figure A represents the original mammogram. The red bounding box in sub-figure B indicates a minimum bounding box approach which captures the mammogram by excluding the noisy labels. Finally, sub-figure C represents the segmented breast region by delineating the pectoral muscle.

4 IMPLEMENTATION DETAILS AND PERFORMANCE EVALUATION

We randomly split individual datasets into 60:20:20 subsets for training, validation, and testing, using a patient-wise splitting protocol, to avoid data leakage issues. We used the training subset to train the models, the validation subset to optimize the hyper-parameters, and the test set for the final model evaluation. VinDr-mammo and mini-DDSM datasets induce class imbalance, and we employed a weighted random sampling approach to handle the class imbalance. (Abd Elrahman & Abraham, 2013).

Experiments were performed on a machine equipped with an Nvidia Tesla V100 16GB graphic card on an Intel Xeon processor provided by the IT Service Centre for Science (CSC) Finland (csc), and in python 3.8 using PyTorch 1.12 (Paszke et al., 2019), as the DL framework enabled with cuda 11.3. We trained the models for 100 epochs using an Adam optimizer with an initial learning rate of 1e-5 at a mini-batch size of 16. The pre-trained CNN classification layer was fine-tuned based on each task, using cross-entropy as the loss function. We evaluated the performance of the trained models on individual test sets with precision, recall, and F1-score, as the evaluation metrics.

5 RESULTS

5.1 PRE-TRAINED CNN MODELS GIVE INCONSISTENT PERFORMANCE FOR DOWNSTREAM CLASSIFICATION TASKS

We investigated different variants (in terms of depth) of five CNN models for mammogram feature extraction. We used these extracted features to train the models for mammogram density assessment and BI-RADS findings classification tasks. Table 2 reveals the pre-trained models' performances for the two downstream classification tasks on the test sets of VinDr-mammo and mini-DDSM datasets with the best numbers bolded for each task. Note that we have investigated different versions of each pre-trained models and only reported the version with higher performance (See B for all the pre-trained CNN model performances for the two downstream tasks). For the density assessment task, Inception Resnet-v2 has shown the highest accuracy of 67% and 68% for the VinDr-mammo and mini-DDSM test sets in terms of F1-score, respectively. For the BI-RADS findings task, the VGG13 and Inception Resnet-v2 models outperform other model in the VinDr-mammo dataset, while Densenet-121 gives the best performance in the mini-DDSM dataset. These results indicate that the pre-trained models achieve inconsistent performances in the two downstream classification tasks for the mammogram datasets. From this experiment, we select Inception ResNet-v2 as the backbone architecture to extract mammogram features independently from each view for different fusion strategies.

5.2 DEEP EVIDENTIAL FUSION NETWORK OUTPERFORMS THE SINGLE-VIEW ASSESSMENT OF MAMMOGRAMS

Table 3 illustrates the results of the two downstream classification tasks with different fusion strategies. The performance of the multi-view evidential fusion strategy (last column of Table 3) has surpassed the single-view for mammogram density assessment and BI-RADS findings tasks in both datasets.

The View-specific late-evidential fusion strategy (VS-LEF, Figure 1 D) outperforms other strategies in both tasks in the mammogram datasets. Considering the multi-view assessment, the VS-LEF strategy shows superior performance than the VS-FLEF strategy (Figure 1 C) by relative F1-score improvements of 7.5% and 3.19% in the VinDr-mammo test set for the mammogram density assessment and BI-RADS findings classification tasks, respectively. Similarly, the VS-LEF strategy outperforms the VS-FLEF strategy by relative F1-score improvements of 21.62% and 4.48% in the mini-DDSM test set for the mammogram density assessment and BI-RADS findings classification tasks, respectively.

The VS-LEF similarly outperforms the LEF strategy (Figure 1 B) by relative F1-score improvements of 1.18% and 4.3% in the VinDr-mammo tests set; and with relative F1-score improvements of 9.75% and 2.35% in the mini-DDSM test set, for mammogram density assessment and BI-RADS findings classification tasks, respectively.

Table 2: Pre-trained CNN model performances on the test sets of the VinDr-mammo and mini-DDSM datasets for density assessment and BI-RADS finding mammogram tasks.

Task	Dataset	Pre-trained models	Precision	Recall	F1-score
Density assessment	VinDr-mammo	VGG13	0.60	0.60	0.60
		Resnet-18	0.64	0.63	0.63
		Inception v3	0.65	0.64	0.64
		Inception Resnet-v2	0.67	0.67	**0.67**
		Densenet-121	0.65	0.65	0.65
		Efficientnet-B1	0.62	0.61	0.61
	mini-DDSM	VGG13	0.57	0.67	0.55
		Resnet-18	0.59	0.68	0.58
		Inception v3	0.58	0.65	0.63
		Inception Resnet-v2	0.68	0.68	**0.68**
		Densenet-121	0.66	0.67	0.67
		Efficientnet-B1	0.63	0.61	0.61
BI-RADS findings	VinDr-mammo	VGG13	0.93	0.95	**0.94**
		Resnet-18	0.93	0.93	0.93
		Inception v3	0.90	0.91	0.91
		Inception Resnet-v2	0.92	0.92	**0.94**
		DenseNet-121	0.92	0.90	0.91
		Efficientnet-B1	0.92	0.92	0.92
	mini-DDSM	VGG13	0.56	0.54	0.50
		Resnet-18	0.58	0.58	0.58
		Inception v3	0.61	0.58	0.54
		Inception Resnet-v2	0.59	0.59	0.59
		Densenet-121	0.62	0.67	**0.68**
		Efficientnet-B1	0.65	0.65	0.65

The VS-LEF outperforms the FLEF strategy (Figure 1 A) by relative F1-score improvements of 6.17% and 2.10% in the VinDr-mammo tests set; and with relative F1-score improvements of 45.16% and 24.28% in the mini-DDSM test set, for mammogram density assessment and BI-RADS findings classification tasks, respectively.

Table 3: The performance evaluation of the single-view and multi-view fusion strategies on the test sets of the VinDr-mammo and mini-DDSM datasets, considering all the evidential fusion strategies: FLEF: feature-level evidential fusion; VS-FLEF: view-specific feature-level evidential fusion; LEF: late-evidential fusion; VS-LEF: view-specific late-evidential fusion.

Task	Dataset	Fusion strategy	L-CC	L-MLO	R-CC	R-MLO	Multi view
Density assessment	VinDr-mammo	FLEF	0.72	0.72	0.72	0.71	0.81
		VS-FLEF	0.79	0.82	0.87	0.83	0.80
		LEF	0.81	0.82	0.81	0.82	0.85
		VS-LEF	0.83	0.84	0.83	0.84	**0.86**
	mini-DDSM	FLEF	0.61	0.61	0.54	0.54	0.62
		VS-FLEF	0.70	0.70	0.76	0.76	0.74
		LEF	0.85	0.85	0.80	0.80	0.82
		VS-LEF	0.84	0.84	0.85	0.85	**0.90**
BI-RADS findings	VinDr-mammo	FLEF	0.93	0.93	0.93	0.93	0.95
		VS-FLEF	0.91	0.91	0.91	0.91	0.94
		LEF	0.94	0.93	0.91	0.92	0.93
		VS-LEF	0.96	0.97	0.93	0.92	**0.97**
	mini-DDSM	FLEF	0.73	0.73	0.70	0.70	0.70
		VS-FLEF	0.81	0.81	0.87	0.87	0.83
		LEF	0.84	0.84	0.81	0.81	0.85
		VS-LEF	0.82	0.73	0.83	0.82	**0.87**

6 DISCUSSION AND CONCLUSION

In this study, we have shown that the multi-view evidential deep fusion learning is a promising approach for the assessment of mammograms, which could be efficiently deployed in a clinical setting. We investigated different pre-trained CNN models and multi-view fusion strategies using evidential deep learning approach on two mammogram classification tasks, namely mammograms density assessment and BI-RADS findings. We empirically demonstrated our findings using VinDr-mammo and mini-DDSM datasets respectively with 5,000 and 1,975 patients with all the four mammogram views available.

The pre-trained CNN models showed inconsistent performances on each task. Appendix B demonstrates the performance of all the pre-trained CNN model variants considered in this study for density assessment and BI-RADS findings tasks. The Inception Resnet-v2 pre-trained model has shown consistent performance for the two datasets and in the two tasks. We additionally interpreted the pre-trained CNN models' classification confidence by visualizing the gradients of the output layer using the saliency maps, illustrated in Figures 4 and 5. Interestingly, the networks with higher accuracies in BI-RADS findings, i.e., the Resnet-101 and Resnet-152, have failed to provide confidence that influences the class probability score for this example. These networks have focused more on the background pixels at the output layer suggesting that intermediate layers might have extracted features, which have been more relevant for the task. Note that transfer learning gap still persists in the medical image domain, specifically in the mammogram-based tasks.

In summary, our experimental results in the two downstream mammogram classification tasks have shown that i) pre-trained CNN models achieve inconsistent performances, ii) the multi-view fusion outperforms the single-view analysis, and iii) the performance of the late-evidential fusion network is superior to the feature-level evidential fusion network.

In future study, we will focus on the following research paths for further improvement; extensive hyper-parameter tuning of the pre-trained CNN models to adapt to the learning of each task, visualizing the intermediate layers' gradients to interpret the model decision at each layer, and finally tuning the evidential loss function to address the class imbalance.

REFERENCES

IT Service Centre for Science (CSC) . URL `www.csc.fi`. Accessed on 25.02.2022.

Shaza M Abd Elrahman and Ajith Abraham. A review of class imbalance problem. *Journal of Network and Innovative Computing*, 1(2013):332–340, 2013.

Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. From generic to specific deep representations for visual recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 36–45, 2015.

Jaime S Cardoso, Inês Domingues, Igor Amaral, Inês Moreira, Pedro Passarinho, João Santa Comba, Ricardo Correia, and Maria J Cardoso. Pectoral muscle detection in mammograms based on polar coordinates and the shortest path. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 4781–4784. IEEE, 2010.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley. Automated mass detection in mammograms using cascaded deep learning and random forests. In *2015 international conference on digital image computing: techniques and applications (DICTA)*, pp. 1–8. IEEE, 2015.

Carl D'Orsi, L Bassett, S Feig, et al. ACR BI-RADS Atlas: Breast Imaging Reporting and Data System. *American College of Radiology, Reston, VA,USA*, 2013.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Krzysztof J Geras, Stacey Wolfson, Yiqiu Shen, Nan Wu, S Kim, Eric Kim, Laura Heacock, Ujas Parikh, Linda Moy, and Kyunghyun Cho. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv preprint arXiv:1703.07047*, 2017.

Naga Raju Gudhe, Hamid Behravan, Mazen Sudah, Hidemi Okuma, Ritva Vanninen, Veli-Matti Kosma, and Arto Mannermaa. Area-based breast percentage density estimation in mammograms using weight-adaptive multitask learning. *Scientific reports*, 12(1):1–19, 2022.

Dianyuan Han. Comparison of commonly used image interpolation methods. In *Conference of the 2nd International Conference on Computer Science and Electronics Engineering (ICCSEE 2013)*, pp. 1556–1559. Atlantis Press, 2013.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. *arXiv preprint arXiv:2102.02051*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Hamza Osman Ilhan, Gorkem Serbes, and Nizamettin Aydin. Decision and feature level fusion of deep features extracted from public covid-19 data-sets. *Applied Intelligence*, 52(8):8551–8571, 2022.

Audun Jøsang. *Subjective logic*, volume 3. Springer, 2016.

Hasan Nasir Khan, Ahmad Raza Shahid, Basit Raza, Amir Hanif Dar, and Hani Alquhayz. Multi-view feature fusion based four views model for mammogram classification using convolutional neural network. *IEEE Access*, 7:165724–165733, 2019.

Charitha Dissanayake Lekamlage, Fabia Afzal, Erik Westerberg, and Abbas Cheddad. Mini-ddsm: Mammography-based automatic age estimation. In *2020 3rd International Conference on Digital Medicine and Image Processing*, pp. 1–6, 2020.

Cheng Li, Jingxu Xu, Qiegen Liu, Yongjin Zhou, Lisha Mou, Zuhui Pu, Yong Xia, Hairong Zheng, and Shanshan Wang. Multi-view mammographic density classification by dilated and attention-guided residual learning. *IEEE/ACM transactions on computational biology and bioinformatics*, 18(3):1003–1013, 2020.

Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *international conference on machine learning*, pp. 7034–7044. PMLR, 2020.

Hieu Trung Nguyen, Ha Quy Nguyen, Hieu Huy Pham, Khanh Lam, Linh Tuan Le, Minh Dao, and Van Vu. Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *medRxiv*, 2022a.

Huyen TX Nguyen, Sam B Tran, Dung B Nguyen, Hieu H Pham, and Ha Q Nguyen. A novel multi-view deep learning approach for bi-rads and density assessment of mammograms. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2144–2148. IEEE, 2022b.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Otávio AB Penatti, Keiller Nogueira, and Jefersson A Dos Santos. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 44–51, 2015.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019.

Rahul Rahaman et al. Uncertainty quantification and deep ensembles. *Advances in Neural Information Processing Systems*, 34:20063–20075, 2021.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Saeed Seyyedi, Margaret J Wong, Debra M Ikeda, and Curtis P Langlotz. Screenet: A multi-view deep convolutional neural network for classification of high-resolution synthetic mammographic screening scans. *arXiv preprint arXiv:2009.08563*, 2020.

Glenn Shafer. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.

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

Rebecca L. Siegel, Kimberly D. Miller, Hannah E. Fuchs, and Ahmedin Jemal. Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, 72(1):7–33, 2022. doi: https://doi.org/10.3322/caac.21708. URL `https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21708`.

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

Tomasz Szandała. Review and comparison of commonly used activation functions for deep neural networks. In *Bio-inspired neurocomputing*, pp. 203–224. Springer, 2021.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Zheng Tong, Philippe Xu, and Thierry Denœux. Fusion of evidential cnn classifiers for image classification. In *International Conference on Belief Functions*, pp. 168–176. Springer, 2021.

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

Jinhua Wang, Xi Yang, Hongmin Cai, Wanchang Tan, Cangzheng Jin, and Li Li. Discrimination of breast cancer with microcalcifications on mammography by deep learning. *Scientific reports*, 6 (1):1–9, 2016.

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

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

A MATHEMATICAL DESCRIPTION OF THE EVIDENTIAL DEEP LEARNING AND THE TRAINING LOSS FUNCTION

Consider the classification task with K mutually exclusive classes, the features that are extracted from the pretrained models are viewed as a sample of Dirichlet distribution with concentration parameters $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_K]$, the probability density function on the vector is given by:

$$Dir(p|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} p_k^{\alpha_k - 1} \tag{3}$$

where, $B(\alpha)$ is the K-dimensional multinomial beta function.

After obtaining the Dirichlet distribution, the subjective logic assigns belief mass $\{b_k\}_k^K$ to each class and an overall uncertainty mass u for the whole classes based on the Dirichlet distribution. The $k+1$ belief masses are all non-negative and their sum is equal to 1:

$$u + \sum_{k}^{K} b_k = 1 \tag{4}$$

where, $u \geq 0$ and $b_k \geq 0$ denote the overall uncertainty and belief mass of the k^{th} class, respectively.

The associated evidence e_k from the input to the classification support, is closely related to the expected concentration parameters α_k , specifically related as $e_k = \alpha_k - 1$. Accordingly, the belief masses are computed as

$$b_k = \frac{e_k}{S}, \quad u = \frac{K}{S} \tag{5}$$

where, $S = \sum_{k=1}^{K}(e_k + 1) = \sum_{k=1}^{K} \alpha_k$ is known as Dirichlet strength. From eq. 5, the more evidence obtained from the k^{th} class, the higher the assigned belief mass is.

A.1 DEMPSTER'S COMBINATION RULE

We employed reduced Dempster's combination rule [15] to combine the evidences from different views. Consider, four mammogram views for K mutually exclusive classes, the reduced Dempster's rule of combination is defined as:

$$M^{\oplus} = \oplus_{v=1}^{V} M^V \tag{6}$$

where, $M^v = \{b_1^v, b_2^v, \ldots, b_k^v, u^v\}$ is the belief mass of the view v. specifically, the fusion of two belief masses $M^1 = \{b_1^1, b_2^1, \ldots, b_k^1, u^1\}$ and $M^2 = \{b_1^2, b_2^2, \ldots, b_k^2, u^2\}$ can be formulated as follows:

$$M^{12} = M^1 \oplus M^2 = \{b_1^{12}, b_2^{12}, \ldots, b_k^{12}, u^{12}\} \tag{7}$$

$b_1^{12} = \frac{1}{1-C}(b_k^1 b_k^2 + b_k^1 u^2 + b_k^2 u^1)$ and $u^{12} = \frac{1}{1-C} u^1 u^2$ where, $C = \sum_{i \neq j} b_i^1 b_j^2$ is the measure of the amount of relative conflict between the two masses. After obtaining the combined belief mass M^{12}, a SoftPlus activation layer [19] is used to output the network class probabilities.

A.2 END TO END MULTI-VIEW FUSION LEARNING FUNCTION

In general, cross-entropy loss function is employed to train a neural network classifier, which is defined as:

$$L_{ce} = -\sum_{j=1}^{K} y_{ij}\ log(p_{ij}) \tag{8}$$

where, p_{ij} is the predicted probability of i^{th} sample for class j. The conventional neural networks can be transformed into evidential neural network by replacing the SoftMax activation with SoftPlus. The output of the i^{th} sample of the SoftPlus activation is parameterized by α_i of the Dirichlet distribution. From the multinomial opinions $D(p|\alpha)$, we get the modified cross-entropy loss function for the evidential neural networks:

$$L_{mce} = \int [\sum_{j=1}^{K} -y_{ij}\ log(p_{ij})] \frac{1}{B(\alpha_i)} \prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1} d(P_i =_{j=1}^{K}\ y_{ij}(\psi(S_j) - \psi(\alpha_{ij}))] \tag{9}$$

where, $\psi(.)$ is the digamma function. To ensure that the correct classified label generates more evidence while the incorrect classified label, low evidence, Kullback libeler divergence term is added as a regularization term to the eq. 9. Thus, the final loss function becomes:

$$L(\alpha_i) = L_{mce} + \lambda_t KL[D(P_i|\alpha_i)||D(p_i|1)] \tag{10}$$

The KL is defined as:
$KL[D(P_i|\alpha_i)||D(p_i|1)] = log(\frac{\Gamma(\sum_{k=1}^{K} \widetilde{\alpha_{ik}})}{\Gamma(K)\prod_{k=1)}^{K}\Gamma(\widetilde{\alpha_{ik}})}) + \sum_{k=1}(\widetilde{\alpha_{jk}} - 1)[\psi(\widetilde{\alpha_{jk}}) - \psi(\sum_{j=1}^{K}(\widetilde{\alpha_{ij}}))]$

where, $\widetilde{\alpha_i} = y_i + (1 - y_i) \bigodot \alpha_i$ is the adjusted Dirichlet distribution to avoid penalizing the evidence, and $\Gamma(.)$ is the gamma function.

The overall loss multi-view evidential loss function is:

$$L_{overall} = \sum_{i=1}^{N}[L(\alpha_i) + \sum_{v=1}^{V} L(\alpha_i^v)] \tag{11}$$

B INCONSISTENT PERFORMANCE OF THE PRE-TRAINED CNN MODELS

Figure 3 illustrates the performance of various pre-trained CNN models for mammogram density assessment and BI-RADS findings classification tasks using mini-DDSM and VinDr-mammo validation sets. For the density assessment task, the VGG19 performs better with 81% in terms of F1-score in the VinDr-mammo evaluation set, while VGG16 exhibits better performance with an F1-score of 70% in the mini-DDSM evaluation set. For the BI-RADS findings task, Inception_Resnet_v2 demonstrates excellent performance with an F1-score of 71% in the VinDr-mammo evaluation set, while Resnet-152 exhibits better performance with an F1-score of 68% in the mini-DDSM evaluation set.

We interpreted the CNN model's classification confidence by visualizing the gradients using saliency maps. Figure **??** illustrates the saliency maps of various pre-trained CNN models for BI-RADS findings tasks on a randomly selected mammogram from the VinDr-mammo evaluation set. Figure 4 shows the benign category mammogram, and the saliency maps demonstrate the necessary pixels that influence the network confidence prediction for each class probability score. The results demonstrate that most CNN models focus on the foreground mammogram pixels. However, the saliency visualizations are not sometimes in agreement with the F1-score values. For instance, in Figure 5, the Inception_Resnet_v2 shows the best performance, however, the saliency visualizations show that the model focuses mostly on the background pixels.

Incorporating the pre-trained CNN models as backbone architecture to extract features is challenging in the medical domain. In future work, we will explore more the pre-trained CNN models by extensive hyper-parameter tuning, loss function generalizations, and by measuring the effect of dataset sizes on the extracted features.

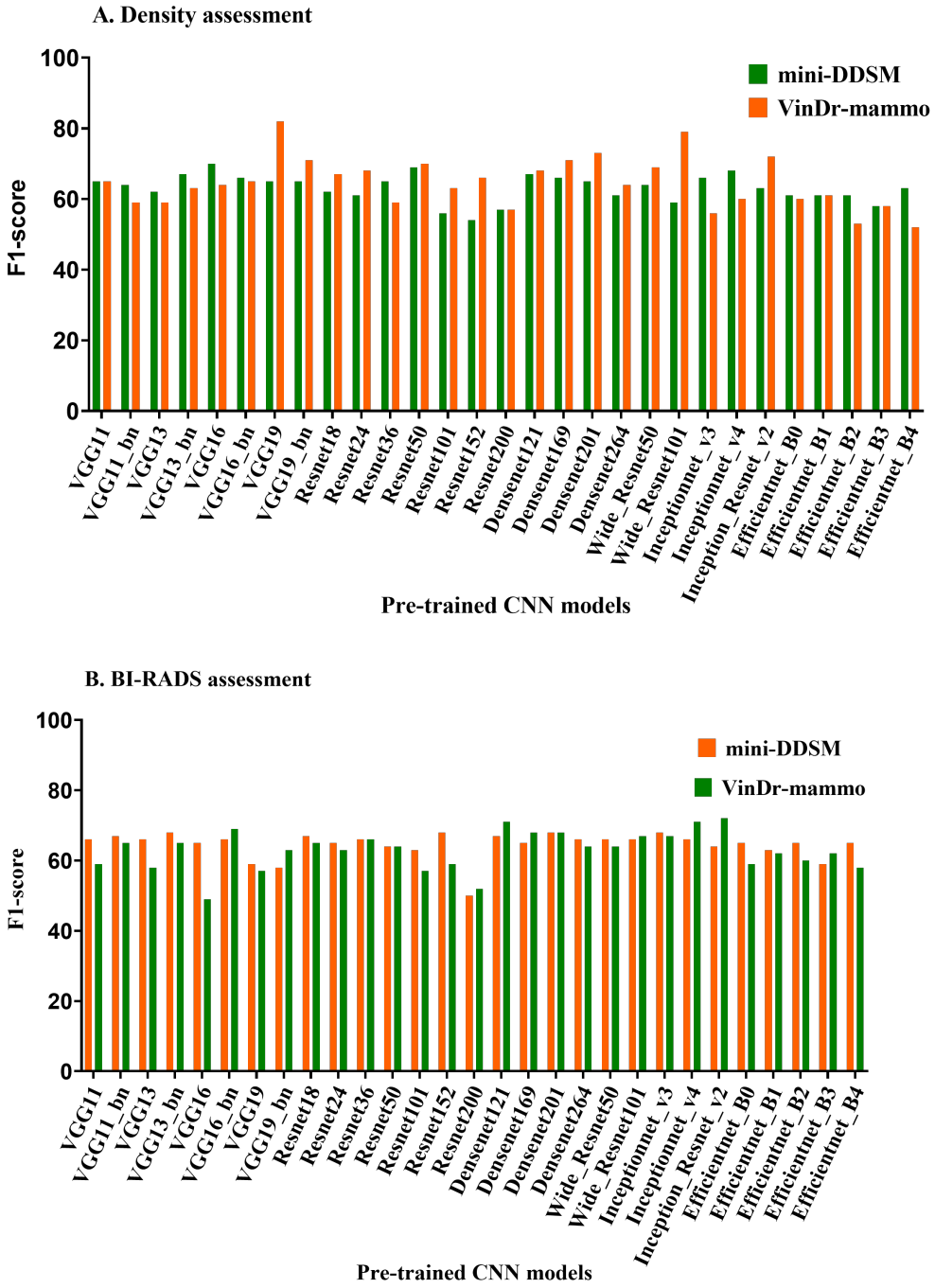

Figure 3: Extensive evaluation of the pre-trained CNN models for the density assessment and BI-RADS findings classification tasks.

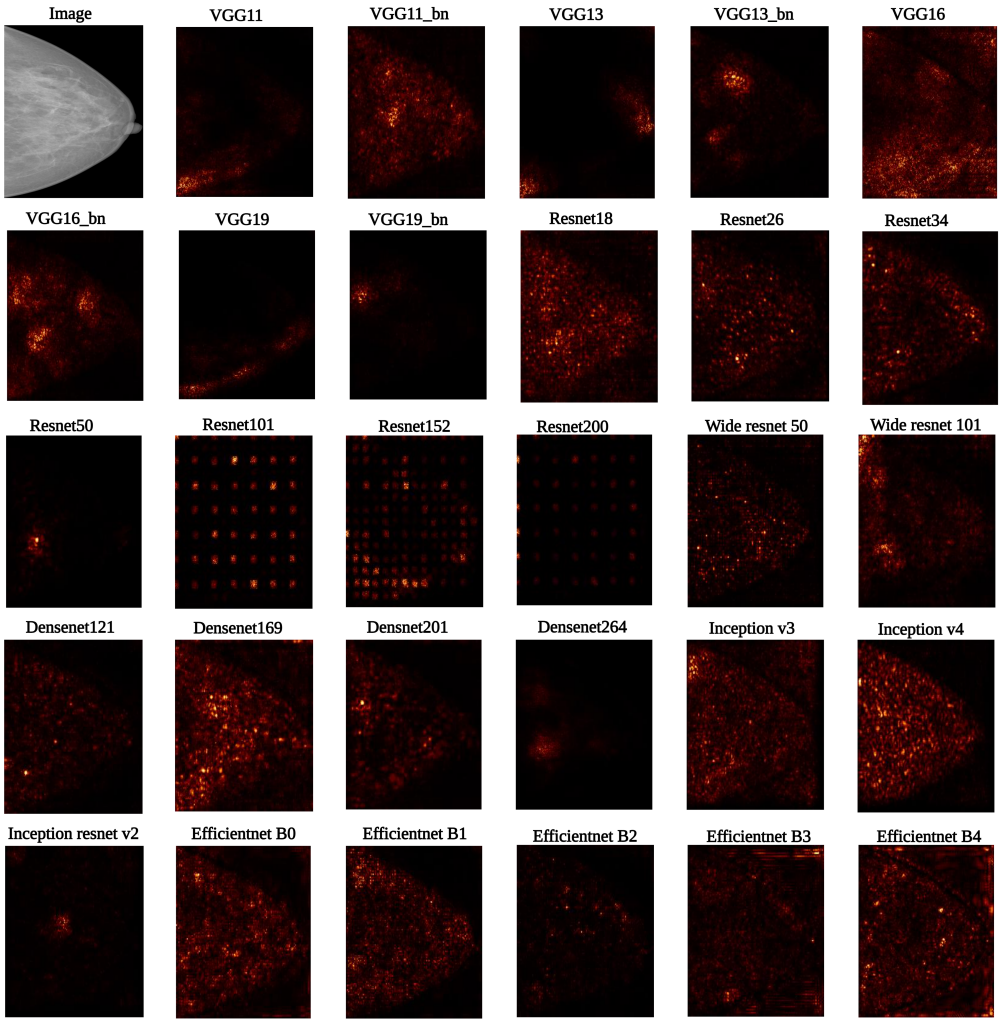

Figure 4: A random benign mammogram from the VinDr-mammo test set. The saliency map visualization indicates the model prediction confidence level. saliency maps show where the model focuses to come up with a decision.

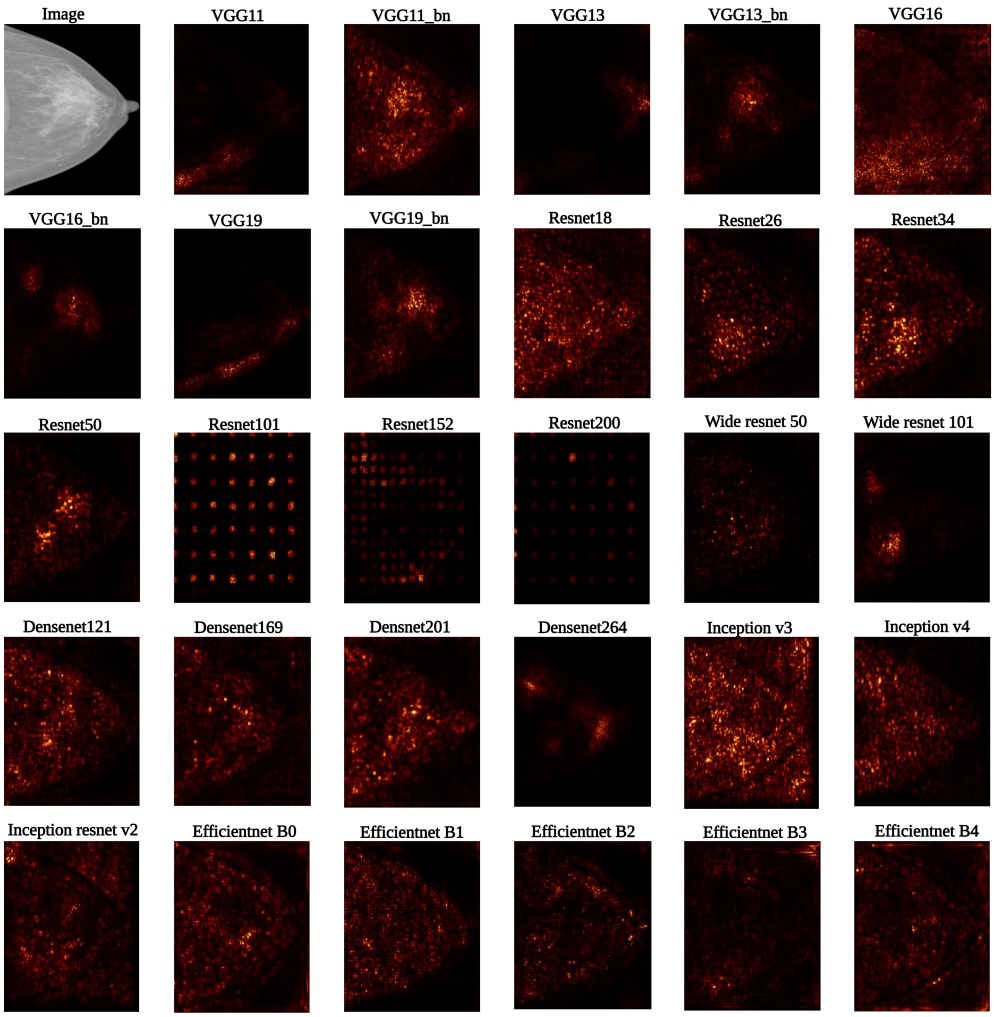

Figure 5: A random benign mammogram from the VinDr-mammo test set. The saliency map visualization indicates the model prediction confidence level. saliency maps show where the model focuses to come up with a decision.