An evidential deep learning framework for assessment of mammograms

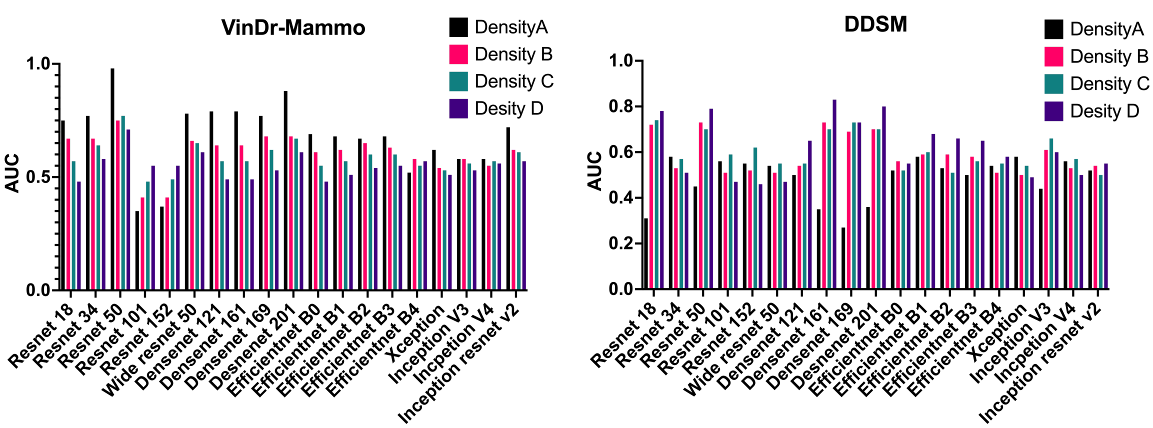
Gudhe Naga Raju[1,*], **Sudah Mazen**[2], **Reijo Sund**[3], **Veli-Matti Kosma**[1,4], **Hamid Behravan**[1,+], **and Arto Mannermaa**[1,4,+]

[1]Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, P.O. Box 1627, Kuopio, 70211, Finland
[2]Department of Clinical Radiology, Kuopio University Hospital, P.O. Box 100, Kuopio, 70029, Finland
[3]Institute of Clinical Medicine, University of Eastern Finland, P.O. Box 1627, Kuopio, 70211, Finland
[4]Biobank of Eastern Finland, Kuopio University Hospital, P.O. Box 100, Kuopio, 70029, Finland
[*]Corresponding. raju.gudhe@uef.fi
[+]The last author position is shared by these authors.

Appendix

A. Compared to other pre-trained CNNs, Resnet50 exhibits consistent performance in the mammogram density assessment task

Table S1 shows that the Resnet50 model has the highest AUCs of 80 $\pm$ 0.024 and 67 $\pm$ 0.04 respectively, on the test sets of VinDr-Mammo and DDSM datasets. Thus, we selected Resnet50 as the preferred backbone architecture for extracting mammogram features for single- and multi-view analyses in this study. Figure S1 further illustrates the performance of different pre-trained CNN models in each density category on the test sets of the VinDr-Mammo and DDSM datasets for the mammogram density assessment task. To address class imbalance during training, we applied random weighted sampling in all models. It is worth noting that the models demonstrated good performance, even for density categories with fewer samples.



(a) VinDr-Mammo test set **(b)** DDSM test set

Figure S1. Illustration of the pre-trained CNN classification model performances on the VinDr-Mammo S1a and DDSM test sets S1b per each density class.

Table S1. Pre-trained CNN model performances on the test sets of the VinDr-Mammo and DDSM datasets for the mammogram density assessment task.

Pre-trained model	**VinDr-Mammo**	**DDSM**
	AUC$\times$100	**AUC**$\times$100
Resnet18	62 $\pm$ 0.03	64 $\pm$ 0.04
Resnet34	66 $\pm$ 0.029	55 $\pm$ 0.05
Resnet50	**80** $\pm$ **0.024**	**67** $\pm$ **0.04**
Resnet101	45 $\pm$ 0.03	53 $\pm$ 0.05
Resnet152	46 $\pm$ 0.03	54 $\pm$ 0.05
Wide Resnet50	68 $\pm$ 0.028	52 $\pm$ 0.05
Densenet121	62 $\pm$ 0.03	56 $\pm$ 0.05
Densenet161	62 $\pm$ 0.03	65 $\pm$ 0.04
Densenet169	65 $\pm$ 0.029	61 $\pm$ 0.04
Densenet201	71 $\pm$ 0.02	64 $\pm$ 0.04
EfficientnetB0	58 $\pm$ 0.03	54 $\pm$ 0.05
EfficientnetB1	59 $\pm$ 0.03	61 $\pm$ 0.04
EfficientnetB2	62 $\pm$ 0.03	57 $\pm$ 0.04
EfficientnetB3	61 $\pm$ 0.03	57 $\pm$ 0.04
EfficientnetB4	55 $\pm$ 0.03	54 $\pm$ 0.05
Inception	55 $\pm$ 0.03	53 $\pm$ 0.05
InceptionV3	56 $\pm$ 0.03	58 $\pm$ 0.04
IncpetionV4	57 $\pm$0.03	54 $\pm$ 0.05
Inception Resnetv2	63 $\pm$ 0.0.29	53 $\pm$ 0.05

B. Mathematical description of the evidential deep learning

Given a dataset $\mathrm{D} = \{\{x_i^v\}_{i=1}^N, y_i^v\}_{v=1}^V$, where N is the total number of training mammogram examinations with V mammogram's views, x_i and y_i represent the i^{th} mammogram and it's corresponding density class label. The backbone CNN architecture Resnet50, parameterized with θ, extracts mammogram's imaging features from each view, $f^v(x_i^v, \theta)$. The extracted features are then passed through a two-layer fully connected neural network with dimensions 512 and 256, with ELU activation and dropout layer in between and softplus in the output layer.

Consider the classification task with K mutually exclusive classes, the output of softplus is viewed as a sample of Dirichlet distribution with concentration parameters $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_K]$, the probability density function is given by[1]:

$$\mathrm{Dir}(p|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} p_k^{\alpha_k - 1}, \tag{S1}$$

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

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

$$u + \sum_{k}^{K} p_k = 1, \tag{S2}$$

where, $u \geq 0$ and $p_k \geq 0$ 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:

$$p_k = \frac{e_k}{S}, \quad u = \frac{K}{S}, \tag{S3}$$

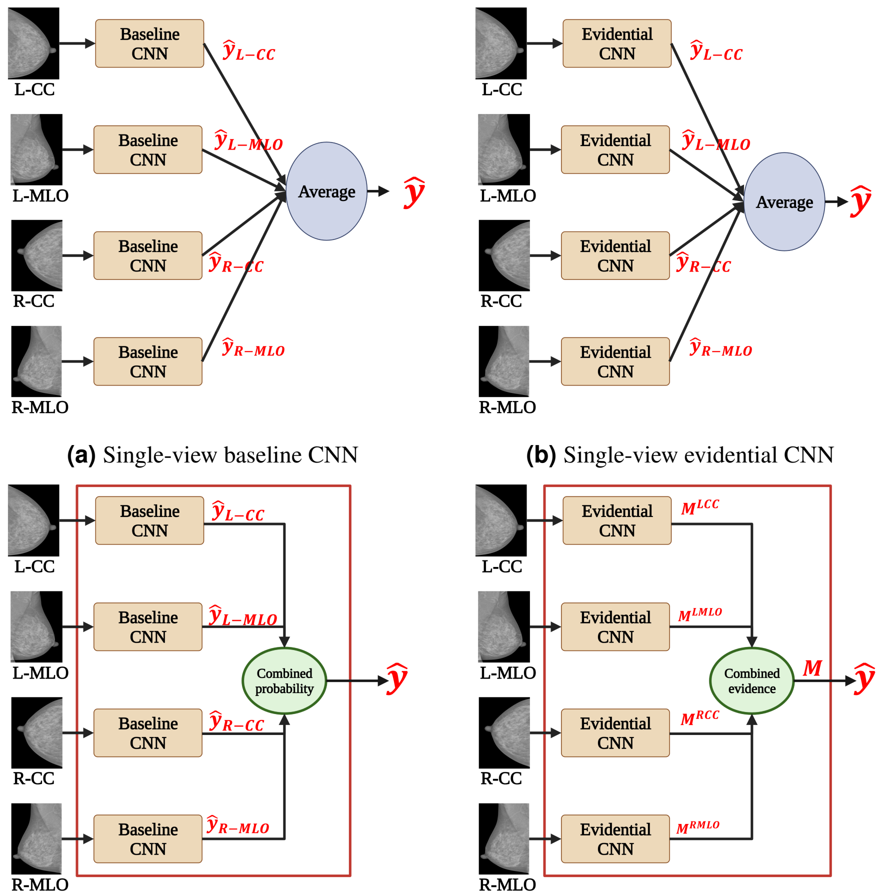
where, $S = \sum_{k=1}^{K}(e_k + 1) = \sum_{k=1}^{K} \alpha_k$ is known as Dirichlet strength. From e.q. S3, the more evidence obtained from the k^{th} class, the higher the assigned belief mass is.

C. Single- and multi-view architectures

We examined both single- and multi-view architectures. For the single-view baseline architecture, as shown in sub-figure S2a, we employed the baseline CNN (sub-figure 1a in main manuscript) on each mammogram's view and then trained each view independently. The final class prediction probability is obtained by averaging the probabilities across all mammogram's views. For the single-view evidential CNN architecture, as shown in sub-figure S2b, we replaced the baseline CNN with the evidential CNN (sub-figure 1b in main manuscript) and then trained each mammogram's view independently. The final density class probability was obtained by averaging the probabilities across all mammogram's views. For the multi-view baseline CNN architecture, as shown in sub-figure S2c, we trained the network end-to-end using the loss function given in e.q. 6 of the main manuscript. The class probability is obtained by computing the average of each mammogram's belief mass. Similarly, for the multi-view evidential CNN architecture, as shown in sub-figure S2d, we replaced the baseline CNN with evidential CNN, and trained end-to-end using e.q. 6 of the main manuscript. With the Dempster's combination rule[2], we combined the belief masses of each mammogram's view.

D. Softplus provides superior performance compared to the softmax non-linear activation function

The results from Table S2 indicate that the MV-DEFEAT (Softplus) model outperforms the MV-DEFEAT (Softmax) model on both datasets for all performance metrics. This is an important finding as it suggests that the MV-DEFEAT (Softplus) model is more effective in detecting patterns and making predictions compared to the MV-DEFEAT (Softmax) model. For the VinDr-Mammo dataset, the MV-DEFEAT (Softplus) model outperforms the MV-DEFEAT (Softmax) model, with a considerably higher AUC (99.5 $\pm$ 0.004). Similarly, for the DDSM dataset, the MV-DEFEAT (Softplus) model outperforms the MV-DEFEAT (Softmax) model, although the difference in performance is smaller compared to the VinDr-Mammo dataset. The MV-DEFEAT (Softplus) model has a higher AUC (88.2 $\pm$ 0.003) compared to the MV-DEFEAT (Softmax) model (85.4 $\pm$ 0.056). The results suggests that the MV-DEFEAT (Softplus) model is better equipped to handle the complexities and nuances of the data, making more accurate predictions.



(a) Single-view baseline CNN

(b) Single-view evidential CNN

(c) Multi-view baseline CNN

(d) MV-DEFEAT

Figure S2. The single- and multi-view architectures investigated in this study. Sub-figure S2a illustrates the single-view baseline CNN architecture, in which each mammogram is processed independently, and the probability decision from each view is then averaged. Sub-figure S2b illustrates the single-view evidential CNN architecture, where the evidence generated from each view is averaged. Sub-figure S2c illustrates the multi-view baseline CNN architecture, where the network is trained end-to-end using multi-view evidential loss function in e.q. 6 of the main manuscript. The class probabilities are computed by considering the average belief masses from each mammogram view. Sub-figure S2d illustrates our proposed approach multi-view evidential CNN architecture (MV-DEFEAT), where the baseline CNN is replaced with the evidential CNN block and the probabilities of all mammogram's views are combined using Dempster's rule of combination.

Table S2. The Performance of the MV-DEFEAT framework utilizing softplus and softmax as non-linear activation functions on the test sets of the VinDr-Mammo and DDSM datasets for the mammogram density assessment task.

Dataset	Model	F1-score	Accuracy	AUC$\times$100	UA
VinDr-Mammo	MV-DEFEAT (Softmax)	89.9 $\pm$ 0.058	85.0 $\pm$ 0.058	82.4 $\pm$ 0.006	77.12 $\pm$ 0.055
	MV-DEFEAT (Softplus)	99.1 $\pm$ 0.008	99.1 $\pm$ 0.008	**99.5** $\pm$ 0.004	87.56 $\pm$ 0.012
DDSM	MV-DEFEAT (Softmax)	69.4 $\pm$ 0.034	69.3 $\pm$ 0.034	85.4 $\pm$ 0.056	59.0 $\pm$ 0.053
	MV-DEFEAT (Softplus)	73.7 $\pm$ 0.004	73.7 $\pm$ 0.004	**88.2** $\pm$ 0.003	64.18 $\pm$ 0.018

E. Improved performance gain with ImageNet pre-trained weights compared to weight initialization

Table S3 compares the performance of MV-DEFEAT when the ResNet50 architecture is trained with the ImageNet weights compared to using Kaiming weight initialization (HE). On the VinDr-Mammo test set, the MV-DEFEAT (ImageNet) model achieved an AUC of 99.9 $\pm$ 0.007, which is considerably higher than the MV-DEFEAT (HE) with AUC of 58.7 $\pm$ 0.058. Similarly, on the DDSM test set, the MV-DEFEAT (ImageNet) achieved an AUC of 86.5 $\pm$ 0.051, which is better than the MV-DEFEAT (HE) model with AUC of 82.4 $\pm$ 0.067. The results indicate that the MV-DEFEAT (ImageNet) model is the superior choice for MV-DEFEAT framework. The improved performance of the MV-DEFEAT (ImageNet) model is likely due to its ability to leverage pre-trained weights from the ImageNet dataset, which contains a large number of images and corresponding labels for many different object classes. This allows the model to better capture important patterns in the data and make more accurate predictions.

Table S3. The Performance of the MV-DEFEAT framework trained with HE and ImageNet weights on the test sets of the VinDr-Mammo and DDSM datasets for the mammogram density assessment task.

Dataset	Model	F1-score	Accuracy	AUC×100	UA
VinDr-Mammo	MV-DEFEAT (HE)	79.1 $\pm$ 0.045	65.9 $\pm$ 0.045	58.7 $\pm$ 0.058	51.18 $\pm$ 0.055
	MV-DEFEAT (ImageNet)	97.6 $\pm$ 0.005	97.7 $\pm$ 0.005	**99.9** $\pm$ 0.007	83.56 $\pm$ 0.012
DDSM	MV-DEFEAT (HE)	58.3 $\pm$ 0.062	58.7 $\pm$ 0.066	82.4 $\pm$ 0.067	60.57 $\pm$ 0.081
	MV-DEFEAT (ImageNet)	66.1 $\pm$ 0.020	67.0 $\pm$ 0.020	**86.5** $\pm$ 0.051	62.10 $\pm$ 0.041

References

1. Bishop, C. M. & Nasrabadi, N. M. *Pattern recognition and machine learning*, vol. 4 (Springer, 2006).

2. Han, Z., Zhang, C., Fu, H. & Zhou, J. T. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis Mach. Intell.* **45**, 2551–2566, DOI: 10.1109/TPAMI.2022.3171983 (2023).