A novel multi-view deep learning approach for BI-RADS and density assessment of mammograms

Huyen T. X. Nguyen[1,†], Sam B. Tran[1,†], Dung B. Nguyen[1], Hieu H. Pham[2,3,*], Ha Q. Nguyen[1,2]

***Abstract*— Advanced deep learning (DL) algorithms may predict the patient's risk of developing breast cancer based on the Breast Imaging Reporting and Data System (BI-RADS) and density standards. Recent studies have suggested that the combination of multi-view analysis improved the overall breast exam classification. In this paper, we propose a novel multi-view DL approach for BI-RADS and density assessment of mammograms. The proposed approach first deploys deep convolutional networks for feature extraction on each view separately. The extracted features are then stacked and fed into a Light Gradient Boosting Machine (LightGBM) classifier to predict BI-RADS and density scores. We conduct extensive experiments on both the internal mammography dataset and the public dataset Digital Database for Screening Mammography (DDSM). The experimental results demonstrate that the proposed approach outperforms the single-view classification approach on two benchmark datasets by huge *F1*-score margins (+5% on the internal dataset and +10% on the DDSM dataset). These results highlight the vital role of combining multi-view information to improve the performance of breast cancer risk prediction.**

***Index Terms*—Mammogram, multi-view deep learning, BI-RADS and density classification.**

I. INTRODUCTION

Breast cancer has now beat lung cancer to become the most commonly diagnosed cancer, according to statistics released by the International Agency for Research on Cancer (IARC) in December 2020 [1]. It is estimated that there were 2.3 million women diagnosed with breast cancer and 685,000 deaths globally in 2020 [2]. Symptoms frequently appear at the later stage of breast cancer, but treatments could be confronted with challenges in this period. Hence, regular breast cancer screening plays a crucial role in the early detection of breast tumors. Mammography, a type of X-ray examination for breasts, is utilized in computer-aided diagnosis (CADx) systems to enhance radiologists' efficiency. A typical mammogram consists of four views: R-CC (right craniocaudal), L-CC (left craniocaudal), R-MLO (right mediolateral oblique), and L-MLO (left mediolateral oblique). In clinical practice, physicians usually use these views for evaluating breast cancer risk. In particular, BI-RADS score is used as a risk assessment and quality assurance tool that supplies a widely accepted lexicon and reporting schema for imaging of the breast [3]. This standard 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). Besides, breast density refers to the amount of fibroglandular tissue in a breast relative to fat; its four descriptors include A (almost entirely fatty breast), B (scattered areas of fibroglandular density), C (heterogeneously dense breast), and D (extremely dense breast) [4].

Previously, many machine learning approaches have been suggested to classify and detect breast cancer using single-view information [5], [6], based on texture descriptors or DL networks. Recently, several studies showed that multi-view approaches [7], [8], [9] improved the diagnosis of breast cancer. The primary technique behind these approaches is building an end-to-end DL model to classify mammograms' pathology. This strategy first extracts features from each view independently and then combines four screening mammography views to produce predictions. In addition, we observed that most current work in breast cancer diagnosis focuses on the discrimination of a mammogram exam as malignant or benign. Nevertheless, few methods could classify mammograms into multi-output. This motivates us to build a DL system that is able to provide a multi-label output, including 5 BI-RADS categories (BI-RADS 1-5) and 4 density classes (A-D). Our experiments show that the proposed multi-view approach consistently outperforms the single-view approach on internal and external benchmark datasets.

Our work makes two main contributions. **First**, we develop a multi-label DL system that can automatically classify the BI-RADS density of mammograms. To our knowledge, we show for the first time in this work that a DL model trained on a large-scale, annotated dataset can predict both BI-RADS and density scores accurately at the same time. **Second**, the proposed approach is based on a novel two-stage multi-view classifier. The first stage learns feature extractors for each view individually. At the second stage, learned features are fused and afterward trained by a LightGBM classifier [10] to predict BI-RADS and density scores. Experimental results show that our multi-view approach outperforms the single-view model up to 5% in terms of *F1*-score. External evaluation on the DDSM dataset [11] also demonstrates the effectiveness of the proposed approach with an increase of 10% in pathology classification compared to the single view approach. The rest of the paper is organized as follows. The proposed approach is presented in Section II. The experiments are provided in Section III. Finally, Section IV discusses the experimental results and concludes the work as well as presents its perspectives.

† The first two authors contributed equally to this work.
* Corresponding author: hieu.ph@vinuni.edu.vn (Hieu Pham).
[1] Smart Health Center, VinBigdata, Hanoi, Vietnam
[2] College of Engineering & Computer Science, VinUniversity, Hanoi, Vietnam
[3] VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam

II. PROPOSED APPROACH

A. Deep Learning-based BI-RADS Density Classification

The proposed BI-RADS density classification architecture has two stages, as illustrated in Figure 1b. The first stage used CNNs to learn latent features from mammograms. The second stage then applied a LightGBM as a classifier to predict BI-RADS and density scores. To verify the effectiveness of the proposed multi-view model, we compare it with a single view approach (Figure 1a).

B. Image Preprocessing

Most mammogram scans have a black background without any information, which is hugely resource-consuming for training. Hence, we built an automatic detector using YOLOv5 [12] to accurately localize the relevant region of the breast from the original scans. To train the detector, we manually annotated 2,000 mammograms, from which 1,600 scans were used for training and 400 scans for validation. The trained model reported an mAP of 0.995 on the validation set and could correctly predict the breast's bounding boxes in unseen mammograms.

C. Multi-view Feature Extraction

According to the feature extraction process, the supervised deep CNNs have been implemented and then removed the fully connected layer at the end of the architecture to get the hidden representations from the input images. The hidden representation is a tensor with dimensions $H \times W \times C$ that is down-sampled from the original dimensions. Specifically, we used ResNet-34 [13] and EfficientNet-B2 [14] for the feature extraction step. To train ResNet-34 extractor, all input images were resized to $H \times W = 512 \times 512$. Additionally, each mammogram has been duplicated from the original grayscale channel to three channels. After passing the trained extractor, we obtained the feature tensor with dimensions of $H \times W \times C = 32 \times 24 \times 512$. Finally, this hidden representation was averaged-pooled over the spatial dimensions to get a 512-dimensional vector. Likewise, to train EfficientNet-B2 model, the input was $H \times W \times C = 1024 \times 768 \times 3$. The hidden tensor was $H \times W \times C = 32 \times 24 \times 1408$ and the hidden vector was in *1408*-dimensional format.

D. BI-RADS & Density Classification

The input of this phase is the feature vectors from the feature extraction step. We trained a LightGBM classifier on $N \times 512$-dimensional vectors provided by ResNet-34 and $N \times 1408$-dimensional vectors provided by EfficientNet-B2, where N is the number of training images. The LightGBM is a gradient boosting framework based on tree algorithms implemented to deliver results faster, reduce memory usage, and improve accuracy. The main idea behind multi-view (Figure 1b) architecture was to use a fusion of multiple hidden information from four views to train an efficient classifier. The four backbones in the feature extraction step are trained for each view of the dataset independently. At this combination technique, we averaged the L-CC and L-MLO feature vectors, and the R-CC and R-MLO feature vectors. Then all the average vectors were used as inputs for LightGBM to provide the confident scores of BI-RADS and density.

III. EXPERIMENTS & RESULTS

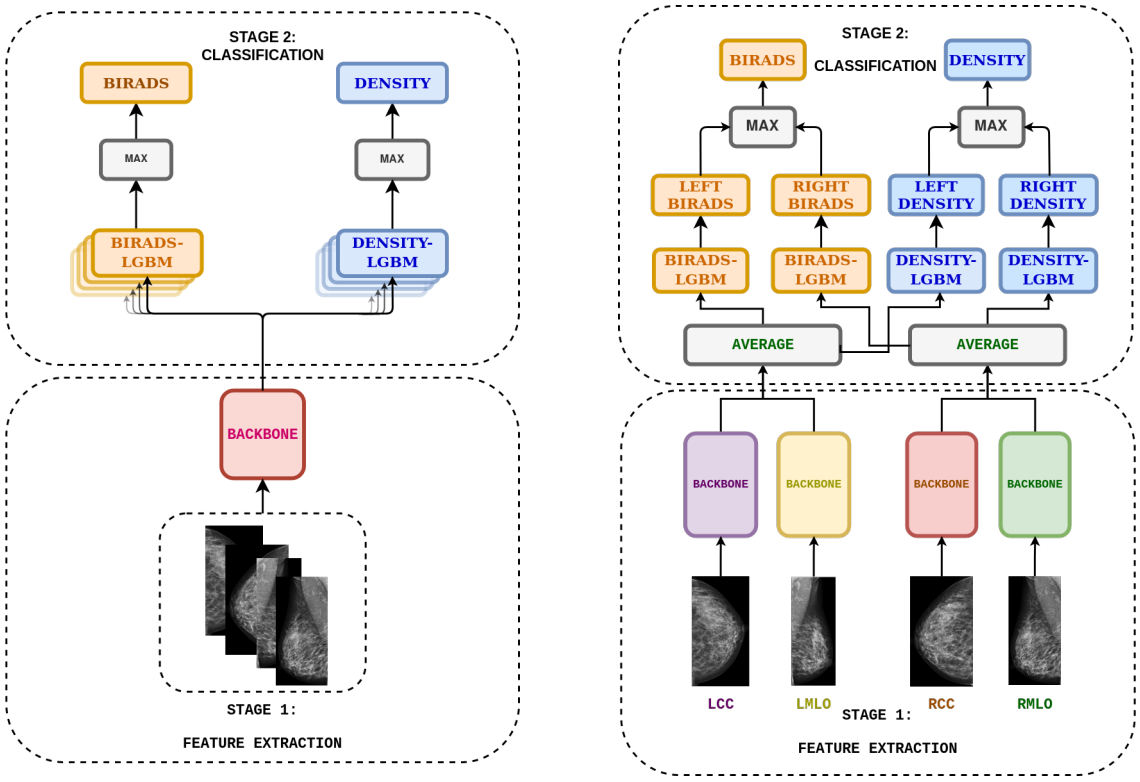
A. Dataset Preparation

We evaluate the proposed method on a part of VinDr-Mammo dataset [15] and DDSM dataset. The private dataset was retrospectively collected from Hanoi Medical University Hospital from 2018 to 2020. Each image was assigned to a team of three radiologists specializing in breast imaging for multilabel annotation (BI-RADS 1-5 and breast density A-D). In total, the dataset includes 36,138 screening mammogram images from 9,911 studies. There are 8509 four-view studies, which consist of four-view images (L-CC, L-MLO, R-CC, R-MLO). These exams were divided into three groups by the multilabel stratification method [16]: training set (5,792 studies), validation set (1,266 studies), and test set (1,271 studies). Descriptions of three sets on the private dataset are provided in Table I. For the DDSM dataset, the number of exams that are multilabel-stratified in training, validation, test set is 1,822; 391; and 391, respectively. Each study involves two-view images of each breast, along with pathology label and density information. There are four types of labels, *normal - benign - cancer - benign with callback - benign without callback*. The DDSM training set contains 486 normal, 704 benign, and 632 cancer cases. The figures of normal/benign/cancer studies in the validation set and test set are 105/147/139 and 104/145/142, respectively. For both datasets, each breast's pathology/BI-RADS label is the maximum label of its CC and MLO images, while the density of these two view images certainly has the same information.

B. Training & Evaluation Metrics

This study was built on PyTorch version 1.8.1 (https://pytorch.org/), and used a machine with an Nvidia GTX 1080 GPU. We trained the feature extractors using SGD optimizer [17] with momentum = 0.9 and cosine annealing learning rate [18]. The cross-entropy function was used to calculate the error. For model evaluation, we used *F1*-score on the 5-class BI-RADS level and 4-class density level. *F1*-score is the harmonic mean of precision and recall. For multi-class problems, macro-*F1* score, which is defined as the mean of class-wise *F1*-scores could be used. As each study's left side and right side might have characteristic differences, the results are appraised for left breasts, right breasts, and study-level for breast diagnosis (BI-RADS types or pathology labels). The evaluation of density cases is based on the left, right, and side-level.

For a fair comparison, we used the same network architecture (ResNet-34/EfficientNet-B2) as the feature extractor with a fixed-size input image for both the single-view and multi-view models. The number of epochs was set to 50, and the training process stopped in case there was no improvement in *F1*-score of the validation set after 15 consecutive epochs by an early stopping callback. The model which acquired the best *F1*-score for validation would be selected



(a) Single-view model (b) Multi-view model (**ours**)

Fig. 1: Illustration of the single view approach (a) and our multi-view approach (b) for BI-RADS and density assessment of mammograms using deep neural networks. Unlike a traditional single-view approach that takes single breast views as inputs during training, the proposed multi-view model takes four breast views as inputs and learns features independently. After extracting hidden features from each view separately, CC and MLO features were combined and fed into a LightGBM classifier to predict the outcomes.

TABLE I: Description of the private dataset used for model development and validation.

Data	Training set					Validation set					Test set				
BI-RADS	LCC	RCC	LMLO	RMLO	Total	LCC	RCC	LMLO	RMLO	Total	LCC	RCC	LMLO	RMLO	Total
1	4,001	4,036	4,006	4,038	16,081	845	871	848	871	3,436	838	852	839	854	3,424
2	1,320	1,337	1,437	1,467	5,561	289	280	287	280	1,136	288	281	286	280	1,135
3	428	416	463	449	1,756	87	76	87	76	326	87	92	87	91	357
4	271	237	729	663	1,900	34	35	34	36	139	31	37	32	38	138
5	104	86	279	223	692	10	4	10	3	27	7	8	7	8	30
Total	6,124	6,112	6,914	6,840	25,990	1,266	1,266	1,266	1,266	5,064	1,271	1,271	1,271	1,271	5,064

as the best feature extractor for the LightGBM classifier. Evaluation metrics of these different network architectures are assessed on the test set.

C. Experimental Results

Table II illustrates the experiment results of two models with different feature extractors (ResNet-34/ EfficientNet-B2) on the private dataset. We observed that the performance of the proposed multi-view model had surpassed the single view model for both BI-RADS and density classification. In the case of BI-RADS classifiers with ResNet-34 backbone, all BI-RADS classes of multi-view model achieve the best *F1*-score compared to single-view. For multi-view architecture, its metric is approximately 6% higher than the single-view model. Similarly, with EfficientNet-B2, the multi-view model impacts more positively to BI-RADS prediction than the single-view model with a *F1*-score of 57.59%. In the case of density categorization, evaluation on the test set of single-view is inferior to the multi-view model. Multi-view results of feature extractor ResNet-34 and EfficientNet-B2, which were calculated on exams, are 56.98% and 61.65% respectively. According to a vast number of executed experiments, the effectiveness of multi-view architecture on our private dataset is proved. Furthermore, we also present the respective performance for two different model architectures on the DDSM dataset. The most important purpose of this experiment is to show the efficiency of the multi-view model on another data distribution. We use EfficientNet-B2 as a backbone to extract the feature representation of DDSM samples. According to our result listed in Table III, our proposed multi-view architecture achieves a higher *F1*-score than single-view by approximately 9.69% at study-level in the classification of pathology and 3.11% at side-level in density classification. The combination of four-view

TABLE II: Quantitative results using *F1*-score of different architectures on the private test dataset.

Backbone	Model	Single-view model			Multi-view model		
ResNet-34	**BI-RADS**	**Left**	**Right**	**Study**	**Left**	**Right**	**Study**
	1	0.78	0.78	0.65	0.85	0.85	0.78
	2	0.52	0.51	0.53	0.53	0.51	0.56
	3	0.17	0.21	0.23	0.16	0.29	0.27
	4	0.13	0.18	0.18	0.14	0.25	0.20
	5	0.62	0.67	0.64	0.73	0.71	0.72
	Macro-*F1*	0.4456	0.4688	**0.4473**	0.4813	0.525	**0.5063**
	Density	Left	Right	Side	Left	Right	Side
	A	0.27	0.12	0.19	0.40	0.33	0.37
	B	0.48	0.54	0.51	0.54	0.62	0.58
	C	0.70	0.71	0.71	0.73	0.74	0.74
	D	0.59	0.61	0.60	0.59	0.60	0.60
	Macro-*F1*	0.5118	0.4973	**0.504**	0.5666	0.5722	**0.5698**
EfficientNet-B2	**BI-RADS**	**Left**	**Right**	**Study**	**Left**	**Right**	**Study**
	1	0.85	0.88	0.80	0.88	0.89	0.82
	2	0.59	0.63	0.63	0.63	0.62	0.66
	3	0.30	0.29	0.35	0.37	0.35	0.43
	4	0.30	0.28	0.28	0.19	0.29	0.28
	5	0.77	0.67	0.72	0.83	0.55	0.70
	Macro-*F1*	0.5617	0.5502	**0.5577**	0.5802	0.5393	**0.5759**
	Density	**Left**	**Right**	**Side**	**Left**	**Right**	**Side**
	A	0.25	0.36	0.32	0.53	0.48	0.50
	B	0.57	0.53	0.55	0.62	0.63	0.62
	C	0.72	0.74	0.73	0.76	0.78	0.77
	D	0.61	0.62	0.61	0.56	0.58	0.57
	Macro-*F1*	0.5386	0.5615	**0.5525**	0.6161	0.6184	**0.6165**

TABLE III: Quantitative results using *F1*-score of EfficientNet-B2 on the DDSM dataset.

Backbone	Model	Single-view model			Multi-view model		
EfficientNet-B2	**Label**	**Left**	**Right**	**Study**	**Left**	**Right**	**Study**
	Normal	0.74	0.70	0.67	0.84	0.77	0.78
	Benign	0.54	0.56	0.44	0.65	0.65	0.62
	Malignant	0.58	0.59	0.61	0.55	0.49	0.61
	Macro-*F1*	0.6186	0.6173	**0.5733**	0.6806	0.6353	**0.6702**
	Density	**Left**	**Right**	**Side**	**Left**	**Right**	**Side**
	A	0.08	0.13	0.11	0.07	0.18	0.13
	B	0.47	0.41	0.44	0.52	0.42	0.48
	C	0.31	0.29	0.30	0.38	0.39	0.39
	D	0.46	0.43	0.45	0.42	0.42	0.42
	Macro-*F1*	0.3319	0.3145	**0.3233**	0.3472	0.3550	**0.3544**

information in mammography can surpass the performance of several variant models of just one view as demonstrated by the results in Table II, III. In conclusion, this multi-view approach significantly impacts BI-RADS and density classification. Description in Section III-A shows that the pathology balance of DDSM dataset might be better than the private dataset. As a result, the improvement score between multi-view model and single-view model on DDSM dataset is superior on pathology level.

IV. DISCUSSION AND CONCLUSION

We introduced in this paper a novel multi-view strategy to classify breast density and estimate BI-RADS scores from mammogram exams. We demonstrated empirically on both the private and public DDSM datasets that the combined information from MLO and CC views improves diagnostic accuracy compared to a single view approach. This opens a new promising approach for building and developing an effective model in early breast cancer detection. We are currently expanding this study by investigating new feature combination strategies to generate more discriminative features for BI-RADS and density classification tasks.

COMPLIANCE WITH ETHICAL STANDARDS

Our work follows all applicable ethical research standards and laws. The study has been reviewed and approved by the institutional review board (IRB) of the hospital. The need for obtaining informed patient consent was waived because this work did not impact clinical care.

ACKNOWLEDGMENT

This study was supported by Smart Health Center, VinBigdata. We would like to acknowledge Hanoi Medical University Hospital for providing us access to their image databases. In particular, we thank all of our radiologists who participated in this project.

REFERENCES

[1] World Health Organization, "Latest global cancer data: Cancer burden rises to 19.3 million new cases and 10.0 million cancer deaths in 2020," International Agency for Research on Cancer, Lyon, France, Dec. 15, 2020 [Online].

[2] World Health Organization, "Breast Cancer, "https://www.who.int/news-room/fact-sheets/detail/breast-cancer, Accessed Aug. 11, 2021 [Online]

[3] Dr Daniel J Bell and Dr Yuranga Weerakkody et al., "Breast imaging-reporting and data system (BI-RADS)," https://radiopaedia.org/articles/breast-imaging-reporting-and-data-system-bi-rads, Accessed Aug. 11, 2021 [Online].

[4] Bruno Di Muzio and Radswiki et al., "Breast Density," https://radiopaedia.org/articles/breast-density, Accessed Aug. 11, 2021 [Online].

[5] Gregoris Liasis, Constantinos Pattichis, and Styliani Petroudi, "Combination of different texture features for mammographic breast density classification," in *International Conference on Bioinformatics Bioengineering (BIBE)*, 2012, pp. 732–737.

[6] Peng Shi, Chongshu Wu, Jing Zhong, and Hui Wang, "Deep learning from small dataset for BI-RADS density classification of mammography images," in *International Conference on Information Technology in Medicine and Education (ITME)*, 2019, pp. 102–109.

[7] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzebski, Thibault Fevry, Joe Katsnelson, Eric Kim, Stacey Wolfson, Ujas Parikh, Sushma Gaddam, Leng Leng Young Lin, Kara Ho, Joshua D. Weinstein, Beatriu Reig, Yiming Gao, Hildegard Toth, Kristine Pysarenko, Alana Lewin, Jiyon Lee, Krystal Airola, Eralda Mema, Stephanie Chung, Esther Hwang, Naziya Samreen, S. Gene Kim, Laura Heacock, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras, "Deep neural networks improve radiologists' performance in breast cancer screening," *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, pp. 1184–1194, 2020.

[8] Hasan Nasir Khan, Ahmad Raza Shahid, Basit Raza, Amir Hanif Dar, and Hani Alquhayz, "Multiview feature fusion based four views model for mammogram classification using convolutional neural network," *IEEE Access*, vol. 7, pp. 165724–165733, 2019.

[9] Krzysztof J. Geras, Stacey Wolfson, S. Gene Kim, Linda Moy, and Kyunghyun Cho, "High-resolution breast cancer screening with multi-view deep convolutional neural networks," *ArXiv*, vol. abs/1703.07047, 2017.

[10] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Q. Ye, and Tie-Yan Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Conference on Neural Information Processing Systems (NIPS)*, 2017, vol. 30, pp. 3146–3154.

[11] Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore, and W. Philip Kegelmeyer, "The digital database for screening mammography," in *Proceedings of the Fifth International Workshop on Digital Mammography*. Medical Physics Publishing, 2001, vol. ISBN 1-930524-00-5, pp. 212–218.

[12] GlennJocher, "Yolov5," https://github.com/ultralytics/yolov5, 2020.

[13] 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*, 2016, pp. 770–778.

[14] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning (ICML)*, 2019, pp. 6105–6114.

[15] Nguyen, Hieu T and Nguyen, Ha Q and Pham, Hieu H and Lam, Khanh and Le, Linh T and Dao, Minh and Vu, Van, "VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography," *arXiv preprint arXiv:2203.11205*, 2022.

[16] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas,"On the stratification of multi-label data," in *Machine Learning and Knowledge Discovery in Databases*, Berlin, Heidelberg, 2011, pp. 145–158, Springer Berlin Heidelberg.

[17] Sebastian Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[18] Ilya Loshchilov and Frank Hutter, "SGDR: Stochastic gradient descent with warm restarts," *International Conference on Learning Representations*, 2017.