Implementation of BI-RADS Classification and Priority Prediction for Mammogram Pre-screening based on Multi-decision Framework

Yi-Chong Zeng*1, Kai-Hsuan Chan*, Yu-Hao Chen*, and Hsin-Yi Lai*
*Digital Transformation Institute, Institute for Information Industry, Taipei, Taiwan, R.O.C.
[1]E-mail: yichongzeng@iii.org.tw Tel: +886-2-66072958

***Abstract*—Mammography is the general way for breast cancer screening worldwide. However, the amount of Negative breast cases is always more than that of Positive ones. It results in radiologists spend much time on the Negative mammograms in the earlier period of screening. This paper introduces a method to assess the risk of breast lesions by two-category BI-RADS. Meanwhile, this method finds out the Positive cases before the Negative ones. First, the detection module finds breast lesions and generates confidence scores to quantify the severity of malignant lesions. Subsequently, the analytics module performs a two-category BI-RADS classifier and then predicts the priority of the mammogram pre-screening based on the multi-decision framework. The experiment results demonstrate that the proposed method achieves higher accuracy of 76% than the compared approaches in two-category BI-RADS classification, and advances the mammogram pre-screening by the 15% number of breast cases at least.**

I. Introduction

Breast cancer is one of the leading illnesses causing women's death worldwide. According to the reports [1], the patient's five-year survival rates in Stage-0 and Stage-1 are approximate to 100% after diagnosing breast cancer. In Stage-4, the five-year survival rate decreases to 22%. In Taiwan, the government offers 0.86 million women for mammogram screening every year. It is helpful to find suspicious abnormal breast tissues and then diagnose breast cancer early. However, breast cancer screening rate over 45-year-old women in 2017 was only 39.8% in Taiwan [2], it was much lower than 70% in the United States [3], 74% in the United Kingdom [4], and 83.1% in Republic of Korea [5].

The general breast imaging is mammography that acquires breast image using a low dose X-ray. Subsequently, approved radiologists screen mammograms and then find suspicious abnormalities, e.g., mass, calcifications, architectural distortion, asymmetry. However, some factors, such as plenty of mammograms, personal skills, and physical fatigue, result in miss-recognition in screening. Moreover, mammogram screening is a time-consuming task. For solving the problems, computer-aided detection (CADe) system was developed for assisting radiologists in mammogram screening. Recently, researchers developed breast lesion detections based on convolutional neural network (CNN), e.g., ResNet, DenseNet, MobileNet, VGG. In [6], Agarwal et al. proposed a method using pre-trained deep learning for mass detection. In [7], Xi et al. used VGG and ResNet for detecting mass and calcifications. For distinguishing malignant tissues from benign ones, Dhunge et al. [8] presented a ResNet-based method, and Gardezi et al. [9] developed a VGG-based method.

American College of Radiology (ACR) designed a quality assurance tool, Breast Imaging-Reporting and Data System (BI-RADS), which standardizes reporting for assessing the risk of developing breast cancer by seven categories [10]. This system is available for mammography, ultrasound, and magnetic resonance imaging (MRI). In general, radiologists assess the BI-RADS of a breast case over multiple mammograms with various view positions and literalities. The final decision-making depends on the highest-risk breast lesion. Therefore, BI-RADS assessment is a challenging task, and it highly relays on radiologist's medical professional skills and experience.

Deep learning was applied to lesion detection and BI-RADS assessment. In [11], Domingues et al. presented a method for classifying BI-RADS of breast cancer using a deep classifier where introduced data augmentation and multi-scale enhancement for performance improvement. In [12], Shen et al. proposed an end-to-end training approach in the deep learning algorithm. They assign images with BI-RADS 1 and 2 as the negative and images with BI-RADS 4, 5, and 6 as the positive. Another BI-RADS assessment depends on text analytics. Castro et al. [13] developed a natural language processing system to extract BI-RADS from radiology reports, and similar work was addressed in [14] by Banerjee et al.

In this paper, we propose a method for improving the performance of mammogram screening, and this method consists of the detection module and the analytics module. First, the detection module performs to breast cases for finding mass and calcifications, which is realized based on CNN-based classifications, such as DenseNet-121 [15] and MobileNet-V2 [16]. The breast case contains four mammograms with two view positions in the left and the right breast. Those view positions are the Mediolateral-Oblique (MLO) and the Cranio-Caudal (CC) views. The proposed detection acquires Class Activation Map (CAM) of an examined mammogram, and then a threshold discriminator applies to the CAM to detect breast lesions.

For the analytics module, we implement two tasks: (1) two-category BI-RADS classification, and (2) priority prediction for mammogram pre-screening. Acquire confidence scores of detected lesions, the proposed method analyzes those scores on a multi-decision framework (MDF) and then determines two-category BI-RADS of a breast case. For two-category BI-RADS, we assign the initial BI-RADS 1 and 2 to the Category-0 (named Negative), and the initial BI-RADS 0, 3, 4, and 5 to the Category-1 (named Positive), which is similar to [12]. Generally speaking, Negative cases are always more than Positive ones. It results in radiologists spent much time on the Negative mammograms in the earlier period of screening. Therefore, the second task is to predict the priority of mammogram screening as well as to find out the Positive cases. The priority prediction analyzes confidence scores of high-risk lesions prediction based on MDF; then, it adjusts the sequence of breast cases according to the sorting of confidence scores. The rest of this paper is organized as follows: the multi-decision framework and the proposed method are introduced in Section II and Section III, respectively. The experiment results are revealed in Section IV, and the conclusion will be drawn in Section V.

II. Multi-decision Framework

The general decision-making approach uses a single classifier, such as Support Vector Machine (SVM). It leads to classification performance depends on a single classifier's ability. In this paper, we develop a decision-making method based on a multi-decision framework (MDF), and it applies to two-category BI-RADS classification and priority prediction for mammogram pre-screening. MDF simulates consensus decision-making so that we regard the output of MDF as the consensus result. Fig.1 is the MDF that consists of multiple decision operators at every stage except for the final one, where $\mathbf{D}^l_i$ denotes the ith decision operator at the lth stage. The inputs of the current decision operator are the outputs of the previous decision operators or the initial data. For example, the inputting data of $\mathbf{D}^3_1$ carries the initial data and the results of $\mathbf{D}^1_7$ and $\mathbf{D}^2_1$. Furthermore, the types of decision operators include a classifier, a regressor, and a voting system. In this paper, our method implements regressions in the decision operators at the first L-1 stages. Those regressions include Support Vector Regression (SVR), Gaussian Process Regression (GPR), and ensemble bagging (BAG).

III. The Proposed Method

Detection module and analytics module are two phases formed the proposed method. In the detection module, detecting breast lesions on mammograms utilizes the existing CNN-based approaches. Subsequently, the analytics module executes two tasks: implementation of two-category BI-RADS classification and priority prediction for mammogram pre-screening. The following sections will introduce the details of those modules.

A. Detection Module

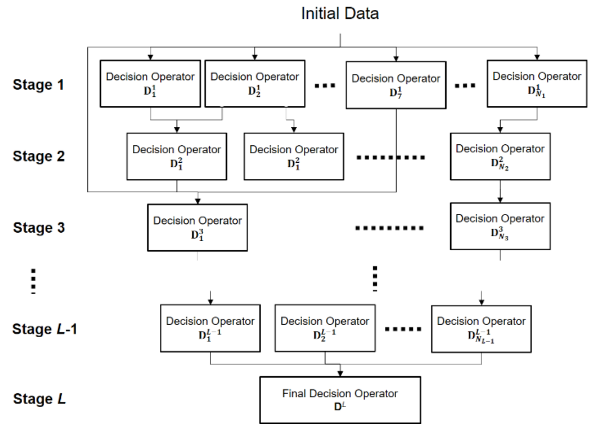


Fig.1. Multi-decision Framework

TABLE I. Types of data for two categories

Category	Mass	Calcification	Two-Category BI-RADS
0	Non-mass	Non-malignant Calcification	Negative (BI-RADS 1 & 2)
1	Mass	Malignant Calcification	Positive (BI-RADS 0, 3, 4 & 5)

In this study, we focus on two kinds of breast lesions, namely mass and calcifications. The proposed detection module is realized based on classification. Fig.2(a) illustrates the general architecture of a CNN-based classifier with two categories. During the training process, inputting data is annotated as either Category-0 or Category-1 and is augmented by mirroring, rotating, scaling, and shifting. TABLE I lists the data types for those two categories. For the mass classifier, Category-0 and Category-1 represent non-mass and mass, respectively. Non-mass include normal tissue and calcifications. For the calcifications classifier, Category-0 and Category-1 represent non-malignant calcifications and malignant calcifications, respectively. Similarly, non-malignant calcifications include normal tissue, mass, and benign calcifications.

Subsequently, a lesion detector is the modification of the trained classifier by removing global averaging pooling, fully connected layer, and Softmax layer. Fig.2(b) displays the architecture of the CNN-based lesion detector. The detector computes Class Activation Map (CAM) of an examined mammogram, such as the two examples in Fig.3. Analyze a mammogram with its CAM, lesions are at the red area that consists of pixels with high CAM values. Therefore, the proposed method executes a thresholding operation on the CAM for detecting lesions. The threshold τ is estimated by,

$$\tau = \underset{t\geq 0}{\arg\max}\left(\sum_{i=1}^{N_0}\Phi\big(S_0(i),t\big)+\sum_{i=1}^{N_1}\Phi\big(S_1(i),t\big)\right) \tag{1}$$

$$\Phi\big(S_j(i),t\big)=\begin{cases}1 & \text{if } S_j(i)<t \text{ and } j=0\\ 1 & \text{if } S_j(i)\geq t \text{ and } j=1\\ 0 & \text{otherwise}\end{cases} \tag{2}$$

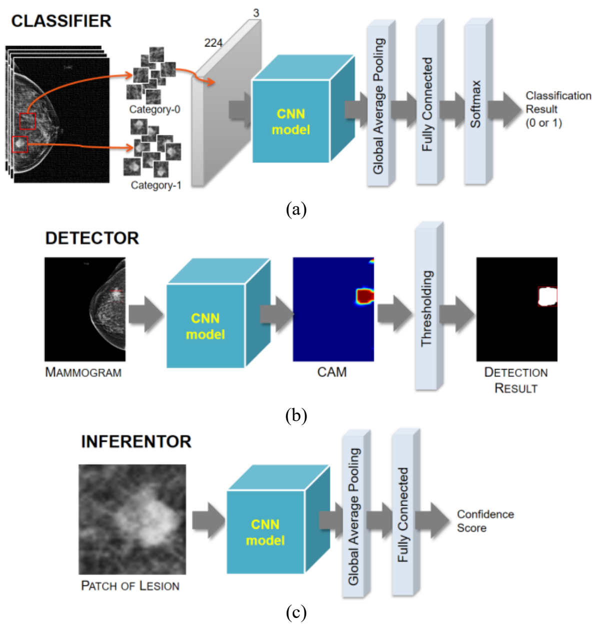
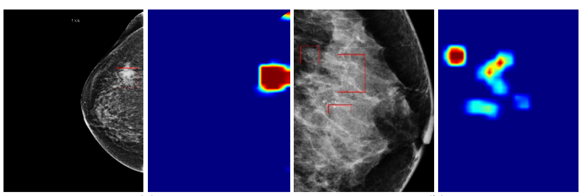


Fig.2. Architectures of (a) CNN-based classification, (b) CNN-based detection, and (c) CNN-based BI-RADS inference



(a) (b)

Fig.3. Examined mammograms and CAMs for (a) mass detection, and (b) calcification detection

where S_j denotes as a set of CAM values with Category-j, and j={0, 1}. The denotations $S_j(i)$ and N_j are, respectively, the ith CAM value and the size of S_j. Finally, our detector retains the pixels whose CAM values are higher than τ, and further groups those pixels as lesions.

The architectures of the mass detector and the first calcifications detector refer to DenseNet-121. Furthermore, our method implements the second calcifications detector using MobileNet-V2. Let Ω_D and Ω_M be the detected areas of calcifications by the first and the second detectors. An ensemble operation intersects Ω_D and Ω_M to generate the final detection result Ω_F, that is $\Omega_F=\Omega_D \cap \Omega_M$. As Ω_D does not overlap Ω_M, both of them are eliminated. The subjective of the ensemble operation is to filter out false-positive lesions as well as to improve detector performance.

B. Analytics Module: Confidence Score Acquiring

After lesion detection, we establish an analytics module to implement two tasks, including two-category BI-RADS classification, priority prediction for breast pre-screening. First, training a primary classifier based on DenseNet-121, Fig.2(a) illustrates the architecture of this classifier. As a lesion is BI-RADS 1 or 2, it is assigned to the Category-0 called Negative case. On the contrary, as a lesion is BI-RADS 0, 3, 4, or 5, it is assigned to the Category-1 called Positive case.

During the training process, inputting data, a set of 224×224-sized lesions, is annotated by either Category-0 or Category-1 in advance. Then, image augmentation applies to the lesions by mirroring, rotating, scaling, and shifting. After training the model, we acquire a BI-RADS inferentor that is the modification of the primary classifier by removing Softmax layer. Fig.2(c) displays the architecture of this inferentor. During the inference process, the inferentor generates a confidence score representing the severity level of an examined lesion. In this work, our method establishes two BI-RADS inferentors of mass and calcifications.

C. Analytics Module: Two-category BI-RADS Classification

In general, radiologist screens four mammograms of a breast case at two view positions in left and right breasts. Those view positions are the craniocaudal (CC) view and the mediolateral oblique (MLO) view. Then, they determine the BI-RADS of the breast case. In this work, our method performs the lesion detector on mammograms in advance. Subsequently, the BI-RADS inferences apply to the detected lesions and generate confidence scores. Finally, find the maximum confidence scores of lesions over four mammograms of a breast case, x_{mass} and x_{calc} are denoted as the maximum confidence scores of the mass and the calcifications, respectively. Here, we aim to establish a two-category BI-RADS classification by analyzing x_{mass} and x_{calc}.

First, x_{mass} and x_{calc} are combined to form a vector x_{comb}, and then there are three types of data. Fig.4 illustrates the block diagram of the proposed two-category BI-RADS classification based on MDF. In the first stage of MDF, the decision operator consists of a regressor and a thresholding discriminator. A voting system applies to the outputs of three thresholding discriminators and generates a final category in the final stage of MDF.

During regressor training, the target of a regressor is the initial two-category BI-RADS. The function of the regressor is to generate a prediction score y_i, where $0 \leq y_i \leq 1$ and $i \in$ {mass, calc, comb}. During BI-RADS inferencing, thresholding discriminator applies to a prediction score y_i and then generates a predicted two-category BI-RADS k_i, which is defined as,

$$k_i = \begin{cases} 1, & \text{if } y_i \geq \tau_i \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where the threshold τ_i is estimated by,

$$\tau_i = \arg\min_{0 \leq t \leq 1} \left((1 - tp_t)^2 + fp_t^2 \right) \tag{4}$$

where tp_t and fp_t denote the true positive rate and the false positive rate. Those values are computed according to the prediction scores under the threshold t, which are defined as follows,

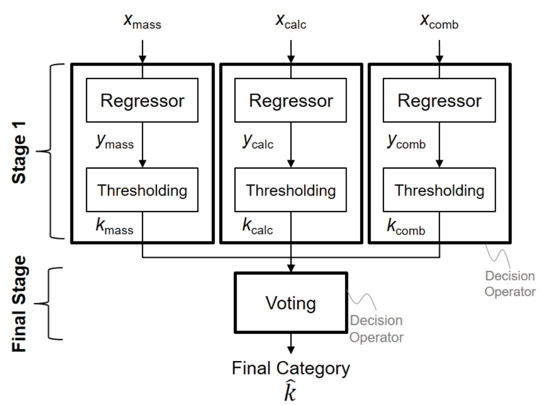


Fig.4. Block diagram of two-category BI-RADS classification based on MDF

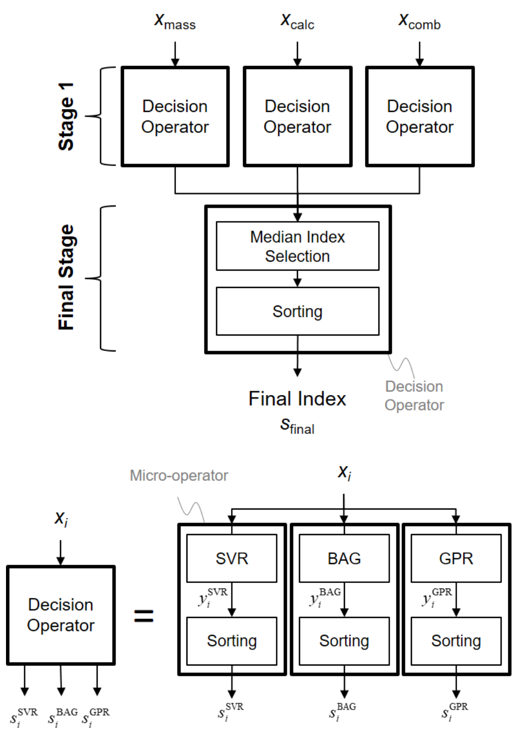


Fig.5. Block diagram of priority prediction for breast pre-screening based on MDF

$$tp_t = \frac{1}{N}\sum_{i=1}^{N}\Psi(y_i, t|1) \tag{5}$$

$$fp_t = \frac{1}{N}\sum_{i=1}^{N}\Psi(y_i, t|0) \tag{6}$$

$$\Psi(y_i, t|\alpha) = \begin{cases} 1, & \text{if } y_i \geq t \text{ and } \alpha = K \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where N is the total number of prediction scores lager than t. The denotation K is the initial two-category BI-RADS, and K={0,1}. Thus, the decision operator is analogous to a classifier. In the final stage, the voting system makes a consensus, and the final category $\hat{k}$ is determined according to,

$$\hat{k} = \begin{cases} 1, & \text{if } \sum_{i=\{\text{mass,calc,comb}\}} k_i > 1 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

In both (3) and (8), the 1-value and the 0-value are, respectively, the Category-0 and the Category-1 of two-category BI-RADS. Those categories are defined in the primary classifier addressed in Section III.B.

D. *Analytics Module: Priority Prediction for Mammogram Pre-screening*

The other task is to predict the priority of a breast case for mammogram pre-screening. In general, the amount of Negative breast cases is more than that of Positive ones. It results in radiologists spent much time on Negative breast cases in the earlier period of screening. For assisting radiologists to improve the performance of screening, the proposed priority prediction finds out Positive cases before Negative ones. Fig.5 illustrates the block diagram of a priority prediction based on MDF.

Three types of data, i.e., x_{mass}, x_{calc}, and x_{comb}, enter three decision operators. Each decision operator is the combination of three micro-operators in the first stage of MDF. The micro-operator further consists of a regressor and a sorting operation. Those three regressions in the decision operator are SVR, BAG, and GPR. During regressor training, the target of a regressor is the initial two-category BI-RADS. The regressor generates a prediction score y^j_i, where $i\in$ {mass, calc, comb} and $j\in$ {SVR, BAG, GPR}. The index s^j_i is derived by sorting y^j_i in descending order. Therefore, a decision operator outputs three indices, and there are nine indices generated in the first stages, i.e., s^{SVR}_{mass}, s^{BAG}_{mass}, s^{GPR}_{mass}, s^{SVR}_{calc}, s^{BAG}_{calc}, s^{GPR}_{calc}, s^{SVR}_{comb}, s^{BAG}_{comb}, s^{GPR}_{comb}. In the final stage of MDF, the decision operator selects the median index s_{med} among those nine indices. The final index s_{final} is acquired by sorting s_{med} of all breast cases in ascending order, and it determines the priority of mammogram pre-screening.

IV. THE EXPERIMENT RESULTS

In this work, we emphasized on two tasks, i.e., two-category BI-RADS classification, and priority prediction for mammogram pre-screening. There were 531 breast cases acquired from the cooperated hospital. Every breast case included four mammograms at two view positions in the left and the right breasts. The details of the experiments and the results are described in the following.

A. *Two-category BI-RADS Classification*

In the first experiment, the proposed method compared with the existing approaches in the two-category BI-RADS classification. First, the lesion detectors introduced in Section

III.A performed to the mammograms for finding mass and calcifications. Subsequently, our BI-RADS inferentor estimated the maximum confidence scores of the detected mass and calcifications. A portion of 531 breast cases was utilized for regressors training (named training data), and the rest cases were the testing data. Let R be the data rate of the total number of the training data to that of the testing data. For each R, we implemented 500 epochs and selected various number of breast cases for training the regressor/classifier in the two-category BI-RADS classification, where R={0.11, 0.25, 0.33, 0.43, 0.5, 0.67, 1, 1.5, 2, 2.33, 4}. Finally, the performance was computed by averaging the accuracies of the 500 epochs, and the accuracy was computed by,

$$accuracy = \frac{1}{M}\sum_{j=1}^{M}\Theta(k_j, K_j) \tag{9}$$

$$\Theta(k_j, K_j) = \begin{cases} 1, & \text{if } k_j = K_j \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where k_j and K_j represent the predicted two-category BI-RADS and the ground truth of the jth case, and M is the number of the testing cases. The proposed two-category BI-RADS classification utilized SVR in the MDF. Furthermore, our method compared to the five existing approaches, including four learning-based approaches and a thresholding one. Those learning-based approaches were SVM, naïve Bayes, random forest, and BAG, which were executed via a single classifier. The thresholding approach determined the two-category BI-RADS according to,

$$k^{\mathrm{TH}} = \begin{cases} 1, & \text{if } x_{\mathrm{mass}} \geq \tau_{\mathrm{mass}} \text{ or } x_{\mathrm{calc}} \geq \tau_{\mathrm{calc}} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where τ_{mass} and τ_{calc} are, respectively, the estimated thresholds for x_{mass} and x_{calc}. Fig.6 illustrates the average accuracies of two-category BI-RADS classification against various data rates by six methods. It shows that our method achieves the highest accuracy over the six methods do, and the best accuracy is 76%. Therefore, the experiment demonstrates that MDF better than a single classifier.

B. *Priority Prediction for Mammogram Pre-screening*

In the second experiment, we focused on the priority prediction of pre-screening by two methods and three states. The ideal state implies that all Positive cases are found before Negative ones. On the contrary, finding all Negative breast cases in advance is the worst state. The initial state is to analyze the initial sequence of breast cases without using any process. The proposed method compares to the max-sorting approach that sorts the breast cases in descending order according to the maximum score x_{mass} and x_{calc}. Similarly, the data rate R controls the numbers of training data and testing data over the 531 breast cases. For each R, we implemented 500 epochs and selected various breast cases for training the regressor in the priority prediction, where R={0.11, 0.25, 0.33, 0.43, 0.5, 0.67, 1, 1.5, 2, 2.33, 4}.

Under the condition of finding a p% number of Positive cases, we record the total number of pre-screening cases M_i of using the ith method, where $0<p<100$, and $i\in${ideal, worst, initial, max-sort, ours}. Then, the reducing rate r_i of the ith method is computed by,

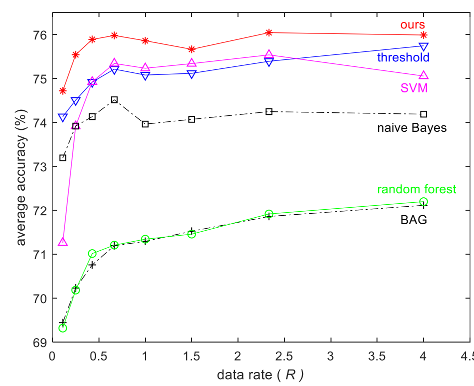


Fig.6. Average accuracies of two-category BI-RADS classification against various data rates by six methods

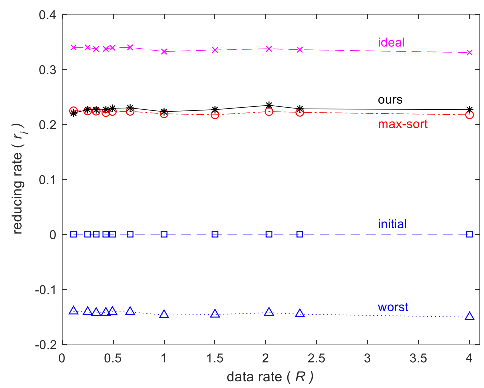


Fig.7. Performance evaluation for priority prediction by two methods and three states against various R under the condition of p=75

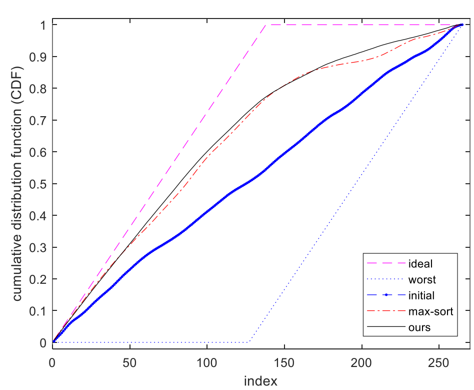


Fig.8. Normalized cumulative curves for finding Positive breast cases by two methods and three states under the condition of R=1

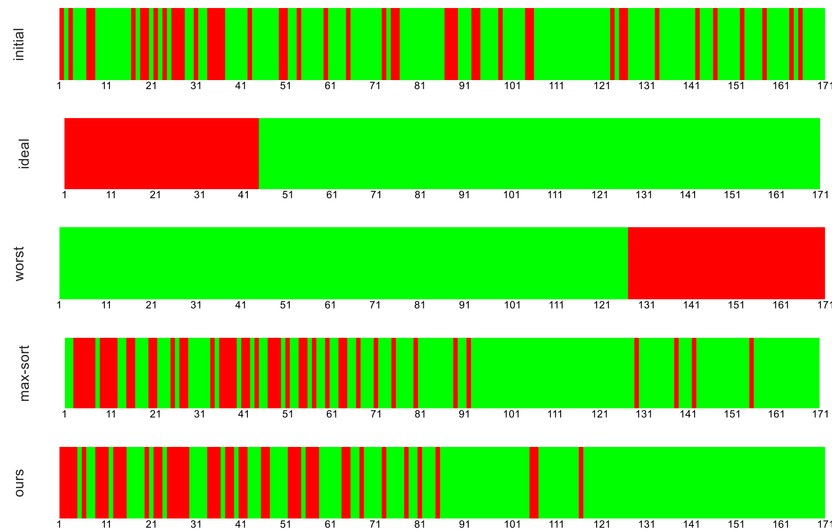


Fig.9. Visualization of priority prediction for two methods and three states (red bar: Positive, green bar: Negative)

$$r_i = \frac{M_{\text{initial}} - M_i}{M} \times 100\% \tag{12}$$

where M is the total number of testing cases, and M_{initial} represents the total number of pre-screening breast cases in the initial state. Fig.7 shows the reducing rates for priority prediction against various data rates under the condition of p=75. Fig.8 illustrates the normalized cumulative curves for finding Positive cases by two methods and three states under the condition of R=1. Therefore, there were 265 testing cases included 138 Positive cases and 127 Negative ones. TABLE II lists the average reducing rates of priority prediction against various p. The experiment result demonstrates that our method achieved better performance than the worst state, the initial state, and the max-sort did. Ours reduced 16%~22% number of pre-screening breast cases compared to the initial state. The problem of the max-sort approach is that two confidence scores (namely x_{mass} and x_{calc}) are not standardized. It results in the sorting process refers to one of the confidence scores and neglects the other. The proposed method generates the consensus by fusing multiple decisions, and it prevents the problem of the max-sort approach depend on one of the confidence scores. Fig.9 displays the visualization of priority prediction for two methods and three states, where the red and the green bars represent Positive and Negative cases, respectively. The number at the bottom of the figures is the index of mammogram pre-screening. While an approach finds Positive cases before Negative ones, it leads to most red bars appear at the left portion of the figure, such as the ideal state. The experiment result demonstrates our method improves the performance of mammogram pre-screening better than the initial state.

TABLE II. Performance evaluation for priority prediction by two methods and three states against various p

Reducing Rate (%)		**p% number of Positive cases**			
		60%	75%	80%	90%
Method	Initial State (Reference)	0	0	0	0
	Ideal State	26.2	33.6	35.6	41.8
	Worst State	–21.9	–14.4	–12.4	–6.3
	Max-sort	18.3	22.2	21.7	8.8
	Ours	19.4	22.7	21.7	16.1

V. CONCLUSION

For assisting radiologists in mammogram screening, this paper introduces a method based on a multi-decision framework to classify the risk of breast lesions into two categories based on BI-RADS. Meanwhile, the proposed method finds out the Positive breast cases before the Negative ones; therefore, it predicts the priority of mammogram pre-screening. The experiment results demonstrate that our method achieves the highest accuracy of 76% over six methods in two-category BI-RADS classification. Furthermore, the proposed priority prediction advances the mammogram pre-screening by reducing 16%~22% number of breast cases.

ACKNOWLEDGMENT

This study is conducted under the "Advanced Artificial Intelligence Technologies and Industry Applications Project (3/4)" of the Institute for Information Industry which is subsidized by the Ministry of Economic Affairs of the Republic of China.

REFERENCES

[1] Timothy Huzar (2018), Life expectancy for stage 3 breast cancer. https://www.medicalnewstoday.com/articles/323177#life-expectancy-and-survival-rates

[2] Utilization of mammography service for women-by age, https://www.gender.ey.gov.tw/gecdb/Stat_Statistics_Query.aspx?sn=p%2fHy1Tg%2bt5fVCNw7Mh1eEQ%3d%3d&statsn=OR0v3mpqY4w1FJ99vq3ytw%3d%3d&d=194q2o4%2botzoYO%2b8OAMYew%3d%3d&n=7952

[3] Breast Cancer Screening (BCS), https://www.ncqa.org/hedis/measures/breast-cancer-screening/

[4] Breast cancer diagnosis and treatment statistics, Cancer Research UK, https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/diagnosis-and-treatment

[5] S. Y. Kang, Y. S. Kim, Z. Kim, H. Y. Kim, H. J. Kim, S. Park, S. Y. Bae, K. H. Yoon, S. B. Lee, S. K. Lee, K.-W. Jung, J. Han, and H. J. Youn, "Breast cancer statistics in Korea in 2017: data from a breast cancer registry," *Journal of Breast Cancer*, vol.23, no.2, pp.115-128, 2020. https://pc.ejbc.kr/Synapse/Data/PDFData/0096JBC/jbc-23-115.pdf

[6] R. Agarwal, O. Diaz, X. Llado, and R. Marti, "Mass detection in mammograms using pre-trained deep learning models," *14th Int'l Workshop on Breast Imaging (IWBI 2018), Int'l Society for Optics and Photonics*, vol. 10718, pp. 107181F, 2018.

[7] P. Xi, C. Shu, and R. Goubran, "Abnormality detection in mammography using deep convolutional neural networks," *2018 IEEE Int'l Symp. on Medical Measurements and Applications (MeMeA)*, pp.1-6, 2018.

[8] N. Dhunge, G. Carneiro, and A. P. Bradley, "Fully automated classification of mammograms using deep residual neural networks," 2017 *IEEE 14th Int'l Symp. on Biomedical Imaging (ISBI 2017)*, pp.310-314, 2017.

[9] S. J. S. Gardezi, M. Awais, I. Faye, and F. Meriaudeau, "Mammogram classification using deep learning features," *2017 IEEE Int'l Conf. on Signal and Image Processing Applications (ICSIPA 2017)*, pp.485-488, 2017.

[10] C. J. D'Orsi, E. A. Sickles, E. B. Mendelson, E. A. Morris, W. E. Creech, P. F. Butler, P. G. Wiegmann, M. B. Chatfield, L. W. Meyer, and P. A. Wilcox, ACR BI-RADS: Breast Imaging Reporting and Data System, Reston, VA, American College of Radiology, 2013.

[11] I. Domingues, P. H. Abreu, and J. Santos, "Bi-RADS classification of breast cancer: a new pre-processing pipeline for deep models training," *IEEE Int'l Conf. on Image Processing (ICIP)*, pp.1378-1382, Oct. 2018.

[12] 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," *Scientific reports*, vol.9, no.12495, pp.1-12, Aug. 2019.

[13] S. M. Castro, E. Tseytlin, O. Medvedeva, K. Mitchell, S. Visweswaran, T. Bekhuis, and R. S. Jacobson, "Automated annotation and classification of BI-RADS assessment from radiology reports," *Journal of biomedical informatics*, vol.69, pp.177-187, April 2017.

[14] I. Banerjee, S. Bozkurt, E. Alkim, H. Sagreiya, A. W. Kurian, and D. L. Rubin, "Automatic inference of BI-RADS final assessment categories from narrative mammography report findings," *Journal of biomedical informatics*, vol. 92, pp.103137-1~103137-11, 2019.

[15] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger," Densely connected convolutional networks," *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.4700-4708, 2017.

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen," MobileNetV2: Inverted Residuals and Linear Bottlenecks," *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510-4520, 2018.