Classification of intrinsic subtypes and histological grade for breast cancers by multimodality images

Chisako Muramatsu, Takumi Iwasaki, Mikinao Oiwa, Tomonori Kawasaki, Hiroshi Fujita

Chisako Muramatsu, Takumi Iwasaki, Mikinao Oiwa, Tomonori Kawasaki, Hiroshi Fujita, "Classification of intrinsic subtypes and histological grade for breast cancers by multimodality images," Proc. SPIE 12286, 16th International Workshop on Breast Imaging (IWBI2022), 122860Y (13 July 2022); doi: 10.1117/12.2625871

SPIE.

Event: Sixteenth International Workshop on Breast Imaging, 2022, Leuven, Belgium

Classification of intrinsic subtypes and histological grade for breast cancers by multimodality images

Chisako Muramatsu[a], Takumi Iwasaki[a], Mikinao Oiwa[b], Tomonori Kawasaki[c], Hiroshi Fujita[d]
[a]Faculty of Data Science, Shiga University,
1-1-1 Banba, Hikone, Shiga 522-8522, Japan
[b]Department of Radiology, Nagoya Medical Center,
4-1-1 Sannomaru, Naka-ku, Nagoya, Aichi 460-0001, Japan
[c]Department of Diagnostic Pathology, Saitama Medical University International Medical Center,
1397-1 Yamane, Hidaka, Saitama 350-1298, Japan
[d]Department of Electrical, Electronics and Computer Engineering, Faculty of Engineering,
Gifu University, 1-1 Yanagido, Gifu 501-1194, Japan

ABSTRACT

Success of breast cancer treatment is subject to various factors, including cancer stage and cancer grade. The best treatment is selected based on the characteristic of cancer. It is desirable to predict the cancer characteristics and prognostic factors accurately and promptly by diagnostic imaging. The purpose of the study is to investigate the use of multimodality diagnostic images in predicting breast cancer subtypes to assist diagnosis and treatment planning. In this study, we classify lesions into molecular subtypes and simultaneously predict histological grades and invasiveness of the cancers by mammography and breast ultrasound images. Models with different architectures including single input and multi-input layers with single head and multiple head models are compared. The results indicate that use of multimodality images is more predictive than using single modalities. The automatic subtype classification using multimodality images may support a prompt treatment planning and proper patient care.

1. PURPOSE

Breast cancer is the most common cancer in women globally [1]. When the cancer is found early, treatment can be highly effective. For successful survival, the best treatment is selected based on the cancer characteristics, including pathologic types, histological grade, and intrinsic subtypes [2-4]. The use of hormonal therapy, chemotherapy, and targeted biological therapy can be determined by the intrinsic subtypes. It may be useful if cancer characteristics can be determined promptly using diagnostic imaging to assist radiologist in tissue sampling and treatment planning.

Several studies have investigated the correlation between molecular subtypes and diagnostic image findings on mammography, ultrasonography, and breast MRI [5-7]. Li et al. [8] proposed quantitative method to classify ER, PR, and HER2 status using radiomic features on MRI and linear discriminant analysis. Zhu et al. [9] investigated the use of deep learning to classify tumors between luminal-A type and others. They reported the use of deep features obtained by the pretrained model and a support vector machine (SVM) provided the better performance than those by end-to-end deep learning models. Ha et al. [10] proposed a subtype prediction method using convolutional neural network (CNN) using MRI data and obtained 70% accuracy on 40 test cases. Zhang et al. [11] compared the use of CNN and a recurrent network (LSTM) for subtype classification on MRI obtained at two centers. Son et al. [12] proposed a machine learning method for prediction of molecular subtypes using radiomic features determined in breast tomosynthesis images. Various features including shape and texture features were extracted based on the manual contours of lesions, and classification between luminal, HER2 and triple-negative types was performed using elastic-net. Ueda et al. [13] investigated the deep learning models in classification of receptor expression status using mammography.

Our purpose in the study is to investigate the image analysis methods to predict molecular subtypes, histological grades, and invasiveness of cancers using mammography and ultrasound images. Subtype classification on diagnostic imaging is a difficult task. Since radiologists generally make diagnostic decisions using multimodality images, which compensate each other, they may be helpful in predicting molecular subtypes. In addition, multitask classification may improve single

16th International Workshop on Breast Imaging (IWBI2022), edited by Hilde Bosmans, Nicholas Marshall, Chantal Van Ongeval, Proc. of SPIE Vol. 12286, 122860Y
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2625871

task classification of subtypes. We compared single modality models and multimodality models with single and double input layers.

2. METHODS

2.1 Database

Digital mammograms and breast ultrasound images used in this study were obtained at Nagoya Medical Center. This study was approved by the institutional review board of Nagoya Medical Center, and informed consent was waived with opportunities for an opt-out. Images were obtained as part of routine screening or diagnostic exams. All the lesions were diagnosed cancer based on biopsy or surgery. Only one lesion per patient was used in this study. The study cases include 346 lesions consisted of 112 luminal-A, 159 luminal-B, 32 HER2, and 43 triple-negative types. The breakdowns of histological grades and invasive or non-invasive cancers are listed in Table 1. Table 2 shows the major image findings of these lesions on mammography. The lesions with no findings include those found by other modalities such as breast tomosynthesis and ultrasound images.

Table 1. Cancer characteristics

Intrinsic subtype	Luminal-A	112
	Luminal-B	159
	HER2	32
	Triple-negative	43
Histological grade	Low	71
	Intermediate	149
	High	126
Invasiveness	Invasive cancer	292
	Non-invasive cancer	54

Table 2. Major image findings on mammography

	Luminal-A	**Luminal-B**	**HER2**	**Triple-negative**
Mass	51	70	12	25
Microcalcifications	14	28	11	4
Architectural distortion	14	24	1	4
Focal asymmetric density	20	17	6	7
No findings	13	20	2	3

Using rough outlines of lesions provided by a radiologist, a square region of interest was cropped from mammograms. For ultrasound images, regions outside the field of view were trimmed. Each patch was resized to 300 x 300 pixels. For ultrasound images, aspect ratio was kept by zero padding.

2.2 Proposed model

We compared models with single modality and multimodality inputs: (1) mammography input, (2) ultrasonography input, (3) mammography and ultrasonography placed side by side as one image, and (4) two modality images separately to two input layers. For each input type, single head model with an output layer of 9 units corresponding to 4 subtypes, 3

histological grades, and 2 pathologic types and multiple head model with three output layers were considered. Figure 1 shows the model architectures. EfficientNetB3 [14] was used as the base model. In the single head model, probabilities for intrinsic subtypes, histological grades, and invasive or noninvasive cancers are determined by a single output layer with sigmoid activation function and binary cross entropy loss. In the multiple head model, subtypes, grades, and pathology types were determined by three softmax layers with categorical cross entropy loss. The number of units in dense/full connection (FC) layer was 128. The optimizer was Adam with a learning rate of 0.001. The models were trained up to 30 epochs.

We also compared the models trained from the scratch, those pretrained with the mass dataset for classification of benign and malignant lesions, and those pretrained with the imagenet. Classification performance was evaluated by 4-fold cross validation with a stratified randomization to balance the classes. In each round, training set was further split into training (80%) and validation (20%) sets, and the model at the epoch that provided the minimum loss for validation set was applied to the test set. Because of the small and unbalanced dataset, oversampling was applied to training set for balancing the classes by random data augmentation. Random shift, scaling, and horizontal flip were applied but image rotation was not used because the data included the ultrasound images.

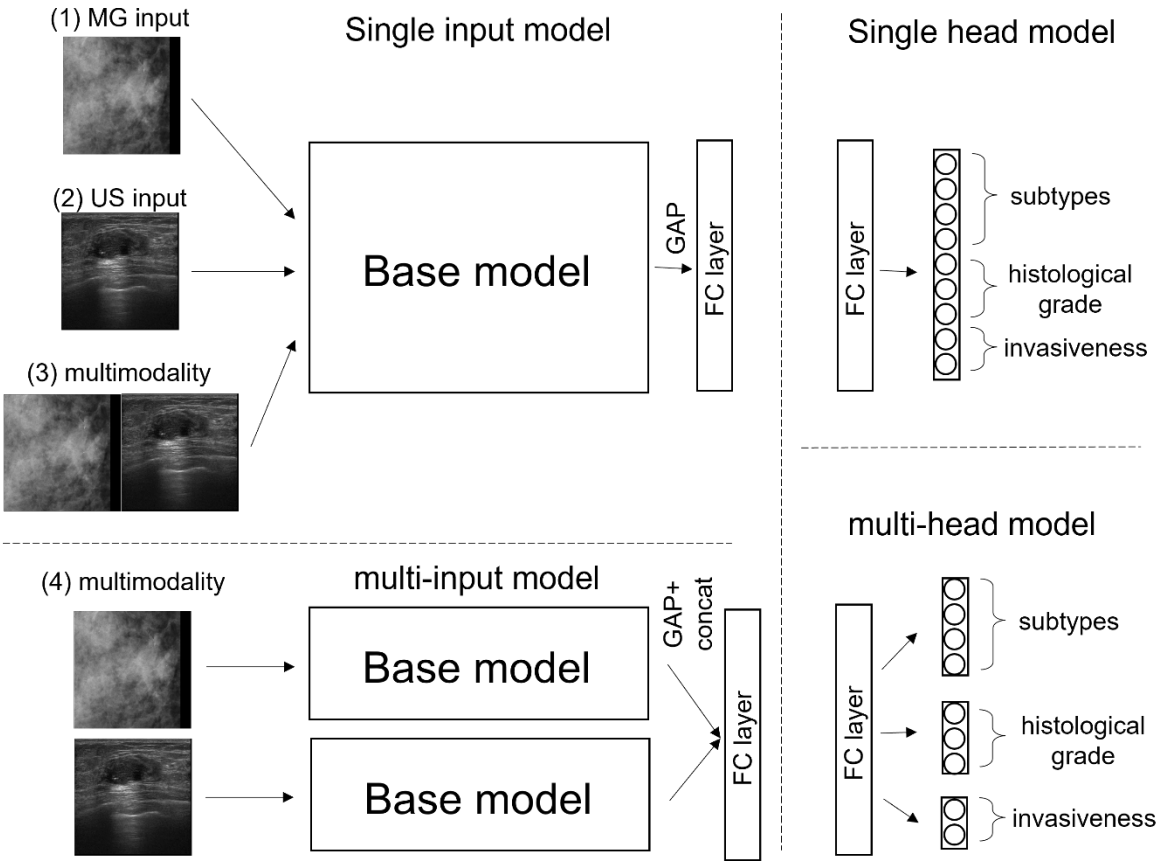

Fig 1. proposed models

2.3 Evaluation

Average (macro) F1 score was determined for comparison of the models. Since the subtypes and invasiveness in our dataset are extremely imbalance, predicting all samples as liminal-A or luminal-B types and as invasive cancer may result in the best accuracies. As in previous studies, performance for 2-class classification as luminal-A versus others, luminal-B versus others, HER2 versus others, and triple-negative versus others are also determined by collapsing 4 classes to 2 classes.

3. RESULTS

The best average F1 scores and the recall rates and accuracies of the corresponding models were 0.32, 0.32, and 0.41 for 4 subtype classification, 0.43, 0.43, and 0.47 for histological grade classification, and 0.59, 0.60, and 0.78 for invasiveness classifications. Table 3 summarizes the results. For all 3 classifications, the two input layer with one head model obtained the best results, although the differences are small in some cases. In general, F1 scores were higher by one-head models than by multi-head models. The best F1 scores using 3 separate models for subtype, grades, and invasiveness classification were also shown in Table 3. The MG model and US model, and 2 input model provided the best score, respectively, for those three. The higher F1 scores were obtained for 3 classifications using multi-task models.

The best F1 scores and the corresponding accuracies for luminal-A versus others, luminal-B versus others, HER2 versus others, and triple-negative versus others were shown in Table 4. For luminal-A versus others, 2 input model provided the best F1 score, whereas MG model obtained the best scores for other three classifications.

Table 5 shows the F1 scores for the model trained from scratch and those using pretrained models. When using the model trained from the scratch and model pretrained with another mass dataset, performances were lower than one pretrained with imagenet dataset. For pretraining with the mass dataset for classification of benign and malignant lesions, 162 benign and 133 malignant images obtained at the same institution but during different time period were used.

Table 3. Classification results for different model architectures

Model		Subtype		Histological grade		Invasiveness	
		F1 score	Accuracy	F1 score	Accuracy	F1 score	Accuracy
(1) MG input	One head	0.319	0.370	0.393	0.442	0.570	0.772
	Multi-head	0.268	0.329	0.319	0.408	0.491	0.815
(2) US input	One head	0.319	0.387	0.366	0.384	0.542	0.751
	Multi-head	0.242	0.376	0.372	0.399	0.575	0.763
(3) Combined image input	One head	0.287	0.396	0.424	0.460	0.569	0.754
	Multi-head	0.236	0.321	0.356	0.471	0.526	0.798
(4) Two inputs	One head	0.**321**	0.410	**0.431**	0.465	**0.590**	0.775
	Multi-head	0.260	0.353	0.344	0.442	0.508	0.653
Separate models		0.280	0.384	0.368	0.401	0.513	0.690

Table 4. Classification results 2-class classification on subtypes

Luminal A vs others		Luminal-B vs others		HER2 vs others		TN versus others	
F1 score	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score	Accuracy
0.604	0.645	0.536	0.540	0.564	0.824	0.579	0.801

4. DISCUSSION

In this study, we attempted to classify breast cancer subtypes using mammography and breast ultrasound images. Classification accuracy was not very high partly due to the fact that it is a difficult task and the dataset is small and imbalance. In addition, the dataset consists of various kind of lesions including masses, microcalcifications, distortions, and those that were originally had no findings on mammograms but found on other modalities. Overall, use of both modality images was effective than using mammography or ultrasound images alone.

Table 5. Classification results for models without and with pretraining

Model	Subtype		Histological grade		Invasiveness	
	F1 score	Accuracy	F1 score	Accuracy	F1 score	Accuracy
No pretraining	0.236	0.379	0.308	0.361	0.529	0.818
Pretrained with mass data	0.305	0.428	0.329	0.361	0.533	0.841
Pretrained with imagenet	0.321	0.410	0.431	0.465	0.590	0.775

The classification performances were slightly improved by training one model for multitask classification than training three separate models. In general, the performances of one-head models were higher than those of multi-head model. The result could be due to the way over-sampling was performed and/or sigmoid activation function. The results require further analysis.

Son et al. [12] proposed a machine learning method to predict molecular subtypes using breast tomosynthesis images. They obtained accuracies of 0.803, 0.704, and 0.507 for classification of triple-negative versus non-triple-negative cases, HER2 versus non-HER2 cases, and luminal versus non-luminal cases, respectively. In terms of accuracy, the results are comparable or slightly better in this study. Their study, however, requires manual outlines of the lesions for determination of hand-crafted features, and the number of validation cases are small (12, 9, and 50 for triple-negative, HER2, and luminal samples). Ueda et al. [13] investigated a deep learning-based method to classify receptor status on mammograms. For 225 test cases, the accuracies for estrogen receptor, progesterone receptor, and HER2 status were 0.63, 0.60, and 0.64, respectively. They obtained the best result by ensemble of different models, such as VGG and inception. Such method may be useful in our study.

In this study, imagenet-pretrained models provided the better performance than model trained from scratch or model pretrained by the mass dataset. Although different task (benign and malignant classification), we conjectured that pretraining with mammographic cases would be effective. However, use of the imagenet dataset was more effective probably because the number of pretraining cases was small. Using the model without pretraining, the output tends to be more biased towards majority cases, whereas outputs were more balanced using the imagenet-pretrained model.

Although performance needs to be improved, this study showed the potential for classifying molecular subtypes, histological grade, and invasiveness of cancer on mammography and breast ultrasound images. Further study is needed with a larger dataset to design useful models for assisting radiologists in breast cancer diagnosis and treatment planning.

ACKNOWLEDGMENT

This study was partly supported by a Grant-in-Aid for Scientific Research (C) (No. 20K08131) by Japan Society for the Promotion of Science.

REFERENCES

[1] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA Cancer J Clin 71, 209-249 (2021).

[2] Perou, C. M., Sorlie, T., Elise, M. B., Rijin van de, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, Hilde., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borrensen-Dale, A. L., Bwown, P. O., Botstein, D., "Molecular portraints of human breast tumours," Nature 406, 747–752 (2000).

[3] Goldhirsch, A., Wood W. C., Coates, A. S., Gelber, R. D., Thurlimann, B., Senn, H. J., "Strategies for subtypes - deeling with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011," Annals Oncol 22, 1736-1747 (2011).

[4] Wolff, A. C, McShane, L. M., Hammond E. aH., Allison, K. H., Fitzgibbons, P., Press, M. F., Harvey, B. E., Mangu, P. B., Bartlett, J. M. S., Hanna, W., Bilous, M., Ellis, I. O., Dowsett, M., Jenkins, R. B., Spears, P. A., Vance, G. H., Viale, G., "Human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical

Oncology/College of American Pathologists Clinical Practice Guideline Focused Update," Arch Pathol Lab Med 142, 1364-1382 (2018).
[5] Uematsu, T., Kasami, M., Yuen, S., "Triple-negative breast cancer: correlation between MR imaging and pathologic findings," Radiology 250, 638-647 (2009).
[6] Celebi, F., Pilance, K. N., Ordu, C., Agacayak, F., Alco, G., Ilgun, S., Sarsenov, D., Erdogan, Z., Ozmen, V., "The role of ultrasonographic findings to predict molecular subtype, histologic grade, and hormone receptor status of breast cancer," Diagn Interv Radiol 21, 448-453 (2015).
[7] Wu, M., Ma, J., "Association between imaging characteristics and different molecular subtypes of breast cancer," Acad Radiol 24, 426-434 (2017).
[8] Li, H., Zhu, Y., Burnside, E. S., Huang, E., Drukker, K., Hoadley, K. A., Fan, C., Conzen, S. D., Zuley, M., Net, J. M., Sutton, E., Whitman, G. J., Morris, E., Perou, C. M., Ji, Y., Giger, M. L., "Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set," npj Breast Cancer 2, 16012 (2016).
[9] Zhu, Z., Albadawy, E., Saha, A., Zhang, J., Harowicz, M. R., Mazurowski, M. A., "Deep learning for identifying radiogenomic associations in breast cancer," Comput Biol Med 109, 85-90 (2019).
[10] Ha, R., Mutasa, S., Karcich, J., Gupta, N., Van Sant, E. P., Nemer, J., Sun, M., Chang, P., Liu, M. Z., Jambawalikar, S., "Predicting breast cancer molecular subtype with MRI dataset utilizing convolutional neural network algorithm," J Digit Imaging 32, 276-282 (2019).
[11] Zhang, Y., Chen J. H., Lin, Y., Chan, S., Zhou, J., Chow, D., Chang, P., Kwong, T., Yeh, D. C., Wang, X., Parajuli, R., Mehta, R. S., Wang, M., Su, M. Y., "Prediction of breast cancer molecular subtypes on DCE-MRI using convolutional neural network with transfer learning between two center," Eur Radiol 31(4), 2559-2567 (2021).
[12] Son, J., Lee, S. E., Kim, E. K., Kim, S., "Prediction of breast cancer molecular subtypes using radiomics signatures of synthetic mammography from digital breast tomosynthesis," Sci Rep 10, 21566 (2020).
[13] Ueda, D., Yamamoto, A., Takashima, T., Onoda, N., Noda, S., Kashiwagi, S., Morisaki, T., Honjo, T., Shimazaki, A., Miki, Y., "Training, validation, and test of deep learning models for classification of receptor expressions in breast cancers from mammograms," JCO Precis Oncol 5, 543-551 (2021).
[14] Tan, M., Le, Q. V., "EfficientNet: rethinking model scaling for convolutional neural networks," arXiv: 1905.11946.