Breast Cancer Early Detection using Preprocessing and Data Enhancement Techniques

Paul Nithya[1,2*] and Nagappan A.[1]
1. Vinayaka Mission's Research Foundation (Deemed to be University), Salem 636308, Tamilnadu, INDIA
2. Federal Institute of Science and Technology, Hormis Nagar, Mookkannoor, Angamaly, Kerala 683577, INDIA
*nithyapaul57@gmail.com

Abstract
Breast cancer is the second most common cause of mortality for women. Early detection and classification of breast cancer is a crucial initial step in its therapy. Different screening methods like MRIs, ultrasounds, mammograms, computed tomography etc. are used to obtain breast images. Because of its capacity to process vast volumes of data, deep learning (DL), a branch of machine learning (ML), has demonstrated impressive outcomes in a number of domains, most notably the biomedical sector.

However, the current deep learning-based breast categorization models have challenges due to the absence of substantial data collection. In order to expand the quantity of images, the proposed method uses a customized generative adversarial network (Cust-GAN) for data augmentation. Additionally, to enhance image quality and remove noise, employ adaptive bilateral filters with weight (ABFW) for image pre-processing.

Keywords: Breast Cancer, Deep Learning, Image Preprocessing Techniques, Data Augmentation, Machine Learning.

Introduction

Breast cancer, which primarily affects women, is one of the illnesses that can afflict both men and women. In order to help doctors diagnose and cure this condition, concerned groups have to learn how to recognize and classify it early on. This dreadful illness has recently spread over the world infecting a large number of women[9,11,21]. The symptoms of breast cancer may include a lump, drainage from the nipple, or changes in the skin's texture around the nipple[1].

There are two types of breast cancer lesions: benign and malignant. Both varieties may be noncancerous or malignant. Based on tissues, hormones and genes, breast cancer is primarily classified into two categories. Both invasive and non-invasive cancers fall under the category of tissues-based cancers. DCIS and lobular are categorized as non-invasive. Paget's illness, invasive lobular carcinoma, invasive ductal carcinoma, inflammatory carcinoma and lymphoma sarcoma are all classified as invasive carcinoma. Breast cancers that are ER-positive or HER2-positive, are categorized as hormone- and gene-based malignancies. Lifesaving measures include raising awareness, detecting problems early and improving therapy continuously[8]. Breast cancer can be detected by screening; it can be diagnosed using ultrasound, magnetic resonance imaging (MRI) and X-ray mammography[14]. Nowadays, a lot of researchers are concentrating on image processing, artificial intelligence, deep learning and machine learning as the field of breast cancer detection research has grown quickly[4].

Researchers are unable to acquire strong classification algorithm performance due to the lack of widely used datasets of breast cancer. Conventional augmentation techniques are severely constrained, particularly when dealing with situations where the images adhere to stringent guidelines like medical datasets. Therefore, researchers use Generative Adversarial Networks (GANs) as a novel way to do data augmentation in addition to the traditional augmentation. By combining conventional and GAN-based augmentation, we were able to attain greater accuracy[2]. In the proposed system, Cust-GAN is used to enrich the input obtained from the dataset. By applying several transformations to the images, Cust-GAN-based image augmentation is used to increase the variety of a training dataset.

Both the quality of the data and the accuracy of machine learning models can be improved by preprocessing datasets for breast cancer diagnosis. Several techniques are used for preprocessing, including understanding the dataset and identifying problems by looking for noise, outliers, missing values, null values and out-of-range values[20]. In the proposed system, the adaptive bilateral filter with weight (ABFW) is significant in preserving image contours and reducing noise more effectively. The objectives of the study are: (i) To increase the dataset's size using the suggested Cust-GAN for data augmentation and (ii) To increase the quality and remove noises in the image by using adaptive bilateral filters with weight (ABFW) for image pre-processing.

Review of Literature

The literature review in line with the suggested system is as follows. Image Preprocessing and augmentation are crucial components of proposed system. The study of Inan et al[13] examines how deep generative models including the conditional generative adversarial network (CTGAN) and the tabular variational autoencoder (TVAE), can help with breast cancer detection and prognosis by producing high-quality synthetic tabular data of breast malignancies. After a thorough analysis, the TVAE model was found to perform better than the synthetic creation of breast tumor data, with

a Chi-Squared test (CS test) score of 0.916 (prognosis) and 0.964 (diagnostic) and a Kolmogorov Smirnov test (KS test) score of 0.887 (prognosis) and 0.928 (diagnosis).

The study introduced a hybrid method for identifying breast tumors from an ultrasound dataset[3]. The U-Net 3+ architecture with GAN to identify blocks for augmentation has successfully increased the segmentation accuracy of breast ultrasound images. According to research of Gab et al[10], there have been problems with tumor classification using MR images including the quantity of pictures in the data set and the low accuracy of the models developed. For data augmentation, PGGAN and traditional augmentation techniques like flipping, rotation and mirroring are employed. Compared to previous models, the VGG19 + CNN model with PGGAN augmentation framework performed more accurately.

Realistic medical data augmentation is essential for training deep learning systems to enhance breast lesion segmentation and classification and to avoid overfitting during the training phase because mammography and Ultra Sound (US) datasets are limited. Four GAN models with various normalization strategies were used in this study[15] to create artificial images. The findings showed that CGAN performed effectively for US data augmentation (FID = 116.03) while SNGAN was effective for mammography data augmentation (FID = 52.89).

Goceri[12] claimed that augmentation techniques based on GANs can boost variety. However, GANs have issues with mode collapse and vanishing gradients. Additionally, if the training process does not guarantee the symmetry and alignment of both the discriminator and generator networks, it is difficult to get adequate training results. Furthermore, GANs are intricate systems, making it challenging to coordinate the discriminator and generator. Combining translation, shearing and rotation is the most effective way to improve the classification of breast mammography pictures as normal, benign and malignant. Abnormal images of the actual medical image size can be produced using the data augmentation technique suggested in the literature[19]. By adding an attention module to both the discriminator and the generator, the quality of the anomalous images produced is significantly improved.

According to Bargsten et al[5], Speckle GAN improves the quality and diversity of generated IVUS images in comparison to a baseline GAN model without a speckle layer. It generates visually appealing images with unique morphology even when trained on minuscule datasets of only 50 images. Nemade et al[18] examined and evaluated the approaches according to their benefits, shortcomings and difficulties as well as the outcomes attained. When compared to other mammography datasets, the majority of the studies that employed the DDSM and MIAS datasets for analysis produced better results. The findings suggest that pattern recognition and parameter selection may benefit from the integration of ML and DL methodologies with optimization strategies. Large, diverse datasets work well with CNN and its variation.

Beeravolu et al[6] offered methods for background removal, pectoral muscle removal, noise addition and image augmentation. The background is successfully removed from 100% of the images using the "Rolling Ball Algorithm" and "Huang's Fuzzy Thresholding.". Using "Canny Edge Detection" and "Hough's Line Transform," 99.06% of the images were found to have no pectoral muscle. Using "Invert," "CTI_RAS," and "ISOCONTOUR" lookup tables (LUTs), image enhancements were used to define the ROIs and the areas inside the ROIs.

For the study report, a preprocessed dataset of breast cancer cases from the real world is used[20]. Breast cancer recurrence prediction without data preprocessing, breast cancer recurrence prediction by error removal and breast cancer recurrence prediction by error removal and filling null values were the three experiments that comprised this case study. These tests were designed to evaluate how preprocessing procedures affected the results of classification algorithms. Breast cancer recurrence prediction models are constructed using three widely used classification algorithms: naïve Bayes, k-nearest-neighbor and sequential minimal optimization. The experiments are assessed using the following metrics: precision, F-measure, accuracy, sensitivity and G-mean. The results show that recurrence prediction is significantly improved by data preprocessing.

This study examines how different data preparation techniques affect the performance of deep learning models[17]. The trial results without data preprocessing revealed that Densenet169, Resnet50 and Resnet101 were the best three models, with accuracy scores of 62%, 68% and 85% respectively. Data augmentation and segmentation helped these models' accuracy rise by 20%, 17% and 6% respectively.

Research Gap: Despite these promising results from the development of the U-Net 3+ architecture employing GAN[3], several problems remain, such as the production of high-quality medical images. In their expanded work, Inan et al[13] chose to investigate how increasing the quantity of synthetically generated data affects the quality of the synthetic data when a variety of generative deep models are merged in different scenarios. Beeravolu et al[6] demonstrated that after processing the mammography pictures, techniques such as Hough's line transform and Canny edge detection are applied. However, its accuracy was only reduced. Therefore, it is critical to employ the best preprocessing and augmentation techniques. The suggested solution aims to get around these current restrictions.

Material and Methods

The proposed model should integrate advanced GAN model for augmentation. The first step should be to gather the

mammography images from real datasets and publicly available datasets like CBIS-DDSM and MIAS. To increase number of images, proposed approach uses customized generative adversarial network (Cust-GAN) for data augmentation. After augmentation, the dataset is to pre-processed using adaptive bilateral filters with weight (ABFW) to increase the quality and remove noises.

Data Acquisition: The publicly accessible CBIS-DDSM[7], MIAS[16] and Real datasets provide the input used to process the suggested model.

Cust-GAN based Data Augmentation: Rotations, flips, translations and colour changes are examples of traditional augmentation techniques. These methods are limited, because they do not add any new information, they just make variations of already images. In order to overcome the issues, Cust-GAN is designed for performing the data augmentation task. Generator and discriminator are the two modules in Cust-GAN, wherein the discriminator attempts to discern between actual and fake images and the generator attempts to produce realistic images in order to trick the discriminator.

As a result, the GAN's adversarial learning model keeps going until the generator creates images that are identical to genuine ones. It assists the augmentation process to obtain the image with high-quality for generating the diverse samples of image. The Cust-GAN is designed based on the Min-Max strategy that utilizes two players in the game Generator and Discriminator. It is formulated as:

$$\min_{A}\max_{B} D(A,B) = \mathrm{E}_{f\sim P_b(f)}[\log B(f)] + \mathrm{E}_{a\sim P_a(a)}[\log(1-B(A(a)))] \quad (1)$$

where the random noise distribution is denoted as $P_a(a)$ and the real data distribution is defined as $P_b(f)$. The outcome of the generator based on the added noise is denoted as $A(a)$ and the detection of real data by the discriminator for the input f is signified as $B(f)$.

Generator network: The generator network in Cust-GAN is designed to generate the high-quality synthesized images. For generating the augmented image, the generator utilizes a deeper network with multiple convolutional layers for enhancing the quality of the generated augmented image. The generator is designed to generate the synthesized image by resizing the image in its initial stage of 128 X 128 X 3. Then, the image is scaled using the tanh activation function and hence the outcome is in the range of [-1,1]. Here, for generating the synthesized image, generator utilized the noise vector of size [100 X 1]. For obtaining the synthesized image with better quality, four convolution layers are utilized in the generator module design.

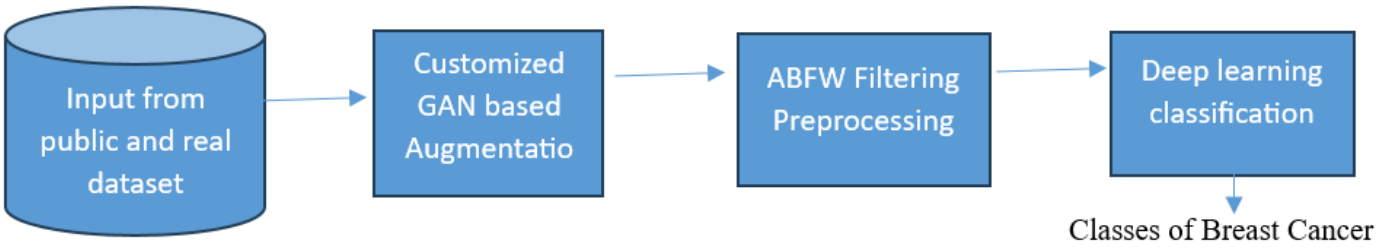


Fig. 1: Work flow of proposed breast cancer classification

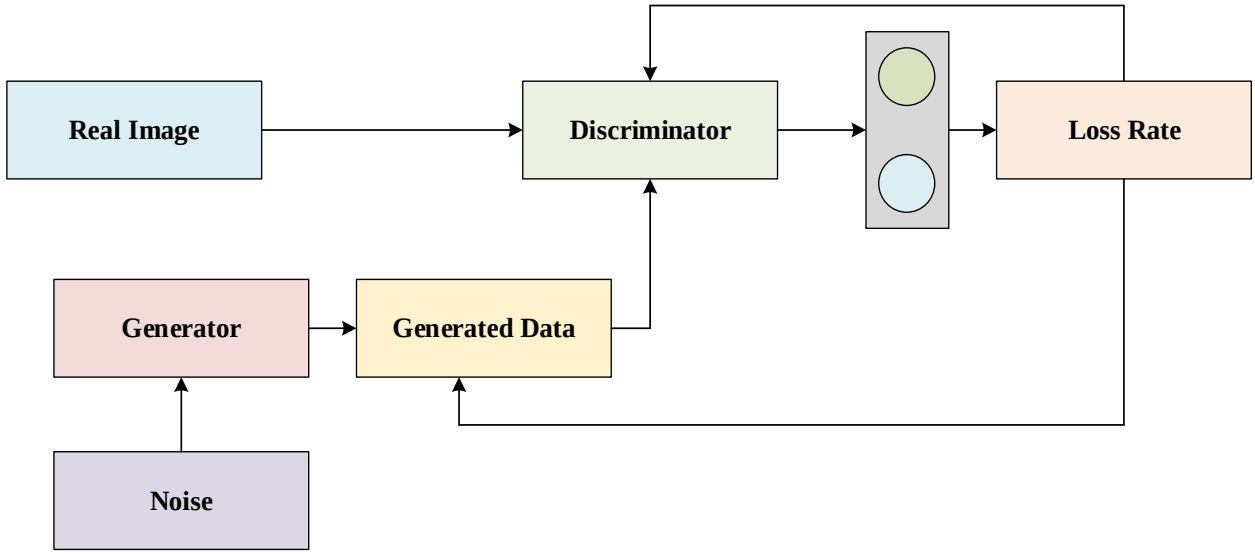


Figure 2: Cust-GAN based Data Augmentation

Discriminator module: The discriminator module of the Cust-GAN is to identify whether the images are real or fake. The discriminator module utilizes an encoder to extract and learn the representation features of the input data. The structure utilized in the discriminator module of the Cust-GAN is portrayed in figure 4.

The feature mapping associated with the discriminator of the Cust-GAN is formulated as:

$$X_{aug}^{pixel} = E_{g\sim B_{encoder}(f), f\sim J_{real}}\left[\left\|\beta(g)-\gamma(f)\right\|\right] \quad (2)$$

where the feature map employed in the discriminator is signified as *g*, wherein the feature map is processed by the function *β*. The function utilized for performing the decoder is defined as $\gamma(f)$ and the real image is denoted by J_{real}.

In the discriminator module, the image resizing is employed initially for obtaining the real part J. Still, the goal of the discriminator module is to accomplish the augmented image *J'* through the efficient feature map criteria. Then, for generating the 128 X 128 image, four various convolution layers are utilized. Finally, the loss function optimization is employed to minimize the information loss during the data augmentation process by matching *J* and *J'*. The loss associated with the generator of the Cust-GAN is formulated as:

$$Loss_B = -E_{f\sim J_{real}}\left[\min\left(0,-1+B(f)\right)\right] - E_{q\sim p(q)}\left[\min\left(0,-1+B(q)\right)\right] + X_{aug}^{pixel} \quad (3)$$

$$Loss_A = -E_{q\sim p(q)}\left[B(A(q))\right] \quad (4)$$

Thus, using the Cust-GAN, the augmented image for the input is accomplished which is fed into the filter for removing the artifacts.

ABFW Filtering: Following data augmentation, the ABFW filtering technique is used to eliminate image artifacts. The same image serves as both the filtering input and the guide when using the traditional bilateral filter. Here, when the image is noisy, the kernel range gets affected that limits the filter's ability to smooth the image effectively. The ABFW filter is significant in preserving image contours and reducing noise more effectively using the separate images for both the guidance and the filtering input. Here, low-pass filtered version of the image is used as the guidance for the range kernel. The outcome of the ABFW, $k(v)$ is expressed as:

$$k(v) = \sum_w G_{ABFW}(v,w)L_w \quad (5)$$

where the weight kernel employed for filtering the image is denoted as $G_{ABFW}(v,w)$ and input is denoted as L_w. The design of weight kernel is defined as:

$$G_{ABFW}(v,w) = \frac{1}{M_v}\exp\left(-\frac{\|v-w\|^2}{2SD^2}\right)\exp\left(-\frac{\|v-\tilde{p}_v\|^2}{2SD^2}\right) \quad (6)$$

where range parameter controlling the intensity similarity is signified as SD and the Gaussian blur is denoted as $\tilde{p}_v$. Then, the expression for the Gaussian blur is formulated as:

$$\tilde{p}_v = \sum_w G_p(v,w)L_w \quad (7)$$

where filter kernel is denoted as $G_p(v,w)$ and is expressed as:

$$G_p(v,w) = \frac{1}{M_v}\exp\left(-\frac{\|v-w\|^2}{2SD_s{}^2}\right) \quad (8)$$

where normalization factor is denoted as M_v and spatial parameter controlling the extent of the blur is signified as $SD_s{}^2$. Thus, by using the ABFW filter, the artifact removal is employed for all the images including the augmented data.

Results and Discussion

The performance of the proposed model is implemented in PYTHON and is analyzed in terms of various assessment measures. These metrics are evaluated by comparing the proposed method with various existing models to show the efficiency and accuracy of the proposed model. The breast cancer classification methods are evaluated based on two various datasets.

CBIS-DDSM dataset[7]: The dataset comprises of mammography images with testing and training data of about 593 images. In this, 286 images are benign and 307 are malignant images. After Cust-GAN based augmentation, the number of samples is increased to 614 malignant and 572 benign images.

MIAS dataset[16]: The MIAS dataset comprises of 321 images with various classes of abnormal category. Asymmetry, architectural distortion, ill-defined masses, spiculated masses, circumscribed masses, normal and calcification are categorized under the malignant type of breast cancer. After Cust-GAN based augmentation, the number of samples is increased to 684 images.

Real Dataset: The real dataset comprises of 124 abnormal images and 257 normal images of mammograms. After Cust-GAN based augmentation, the number of samples is increased to 248 malignant and 514 benign images. The experimental outcome for the samples from the dataset based on the pre-processing is illustrated in figure 5.

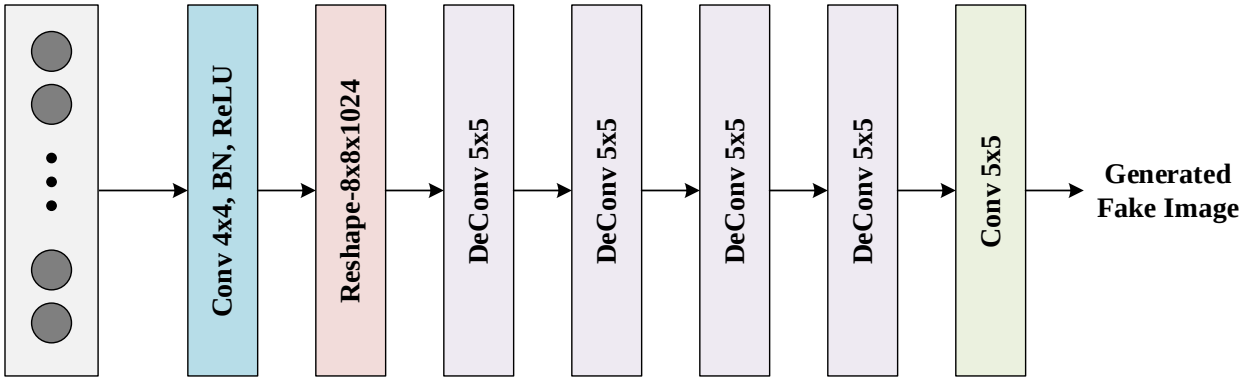


Figure 3: Structure of Generator

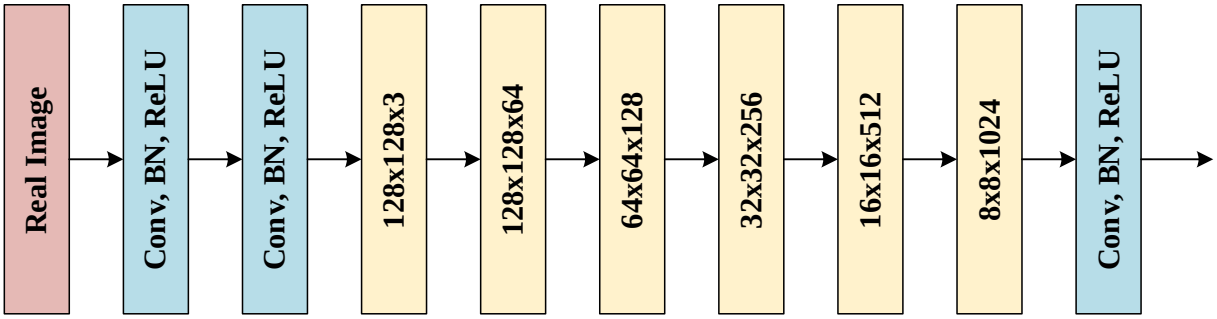


Figure 4: Structure of Discriminator

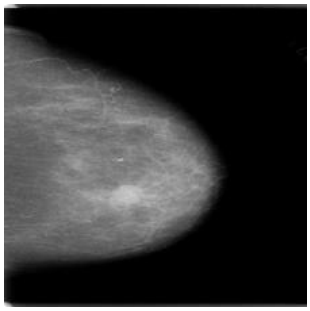
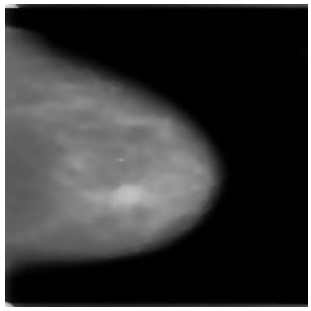
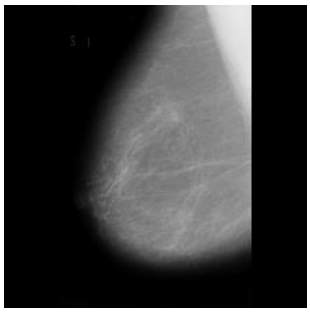
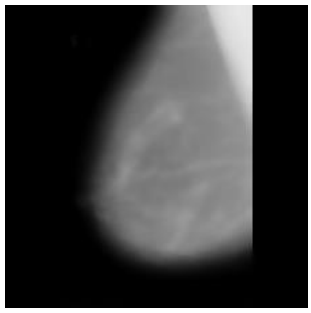
Dataset	Input Image	Pre-Processed Outcome
CBIS-DDSM		
MIAS dataset		
Real dataset		

Figure 5: Experimental outcome of proposed model

Conclusion

This study introduces a deep learning model for the classification of breast cancer in order to aid in early illness diagnosis. The designed model utilized Cust-GAN for augmenting the data more efficiently. In this case, the data augmentation aids in strengthening the deep learning model's generalization and robustness, which raises the classification accuracy. Here, ABFW effectively reduces noise in the images while preserving important edges and details, which are crucial for accurate diagnosis of the disease.

References

1. Abd-Elnaby Muhammed, Marco Alfonse and Mohamed Roushdy, Classification of breast cancer using microarray gene expression data: A survey, *Journal of Biomedical Informatics*, **117**, 103764 **(2021)**

2. Aqthobirrobbany Aqil et al, A systematic review of breast cancer detection on thermal images, *Communications in Science and Technology*, **8(2)**, 216-225 **(2023)**

3. Al-Dhabyani Walid et al, Deep learning approaches for data augmentation and classification of breast masses using ultrasound images, *Int. J. Adv. Comput. Sci. Appl*., **10(5)**, 1-11 **(2019)**

4 Alruily Meshrif et al, Breast Ultrasound Images Augmentation and Segmentation Using GAN with Identity Block and Modified U-Net 3+, *Sensors*, **23(20)**, 8599 **(2023)**

5. Bargsten Lennart and Alexander Schlaefer, SpeckleGAN: a generative adversarial network with an adaptive speckle layer to augment limited training data for ultrasound image processing, *International Journal of Computer Assisted Radiology and Surgery*, **15**, 1427-1436 **(2020)**

6. Beeravolu Abhijith Reddy et al, Preprocessing of breast cancer images to create datasets for deep-CNN, *IEEE Access*, **9**, 33438-33463 **(2021)**

7. CBIS-DDSM dataset: https://www.kaggle.com/datasets/awsaf49/cbis-ddsm-breast-cancer-image-dataset

8. Dubey A.K., Gupta U. and Jain S., Analysis of k-means clustering approach on the breast cancer Wisconsin dataset, *Int J Comput Assist Radiol Surg*., **11(11)**, 2033–2047 **(2016)**

9. Fu Y., Lei Y., Wang T., Curran W.J., Liu T. and Yang X., A review of deep learning - based methods for medical image multi-organ segmentation, *Phys. Med.*, **85**, 107–122 **(2021)**

10. Gab Allah, Ahmed M., Amany M. Sarhan and Nada M. Elshennawy, Classification of brain MRI tumor images based on deep learning PGGAN augmentation, *Diagnostics*, **11(12)**, 2343 **(2021)**

11. Gheshlaghi S.H., Kan C.N.E. and Ye D.H., Breast Cancer Histopathological Image Classification with Adversarial Image Synthesis, In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Virtual, **1–5**, 3387–3390 **(2021)**

12. Goceri Evgin, Medical image data augmentation: techniques, comparisons and interpretations, *Artificial Intelligence Review*, **56(11)**, 12561-12605 **(2023)**

13. Inan Muhammad Sakib Khan, Sohrab Hossain and Mohammed Nazim Uddin, Data augmentation guided breast cancer diagnosis and prognosis using an integrated deep-generative framework based on breast tumor's morphological information, *Informatics in Medicine Unlocked*, **37**, 101171 **(2023)**

14. Iranmakani S., Mortezazadeh T., Sajadian F., Ghaziani M.F., Ghafari A., Khezerloo D. and Musa A.E., A review of various modalities in breast imaging: technical aspects and clinical outcomes, *Egypt J Radiol Nucl Med*, **51**, 1–22 **(2020)**

15. Jiménez-Gaona Yuliana et al, Gan-based data augmentation to improve breast ultrasound and mammography mass classification, *Biomedical Signal Processing and Control*, **94**, 106255 **(2024)**

16. MIAS dataset: https://www.kaggle.com/datasets/kmader/mias-mammography

17. Mohamed Ammar et al, The impact of data processing and ensemble on breast cancer detection using deep learning, *Journal of Computing and Communication*, **1(1)**, 27-37 **(2022)**

18. Nemade Varsha, Pathak Sunil and Dubey Ashutosh Kumar, A systematic literature review of breast cancer diagnosis using machine intelligence techniques, *Archives of Computational Methods in Engineering*, **29(6)**, 4401-4430 **(2022)**

19. Qi Chang et al, SAG-GAN: Semi-supervised attention-guided GANs for data augmentation on medical images, *arXiv preprint arXiv*, DOI:10.48550/arXiv.2011.07534 **(2020)**

20. Sajjadnia Zeinab, Raof Khayami and Mohammad Reza Moosavi, Preprocessing breast cancer data to improve the data quality, diagnosis procedure and medical care services, *Cancer Informatics*, **19**, 1176935120917955 **(2020)**

21. Singh S. and Tripathi B.K., Pneumonia classification using quaternion deep learning, *Multimed. Tools Appl*., **81**, 1743–1764 **(2022)**.

(Received 09th January 2025, accepted 15th March 2025)