# Deep Learning Approaches for Data Augmentation and Classification of Breast Masses using Ultrasound Images

Walid Al-Dhabyani[1], Aly Fahmy[4]
Faculty of Computer and Information,
Cairo University, Cairo, Egypt

Mohammed Gomaa[2], Hussien Khaled[3]
National Cancer Institute,
Cairo University, Cairo, Egypt

***Abstract*—Breast classification and detection using ultrasound imaging is considered a significant step in computer-aided diagnosis systems. Over the previous decades, researchers have proved the opportunities to automate the initial tumor classification and detection. The shortage of popular datasets of ultrasound images of breast cancer prevents researchers from obtaining a good performance of the classification algorithms. Traditional augmentation approaches are firmly limited, especially in tasks where the images follow strict standards, as in the case of medical datasets. Therefore besides the traditional augmentation, we use a new methodology for data augmentation using Generative Adversarial Network (GAN). We achieved higher accuracies by integrating traditional with GAN-based augmentation. This paper uses two breast ultrasound image datasets obtained from two various ultrasound systems. The first dataset is our dataset which was collected from Baheya Hospital for Early Detection and Treatment of Women's Cancer, Cairo (Egypt), we name it (BUSI) referring to Breast Ultrasound Images (BUSI) dataset. It contains 780 images (133 normal, 437 benign and 210 malignant). While the Dataset (B) is obtained from related work and it has 163 images (110 benign and 53 malignant). To overcome the shortage of public datasets in this field, BUSI dataset will be publicly available for researchers. Moreover, in this paper, deep learning approaches are proposed to be used for breast ultrasound classification. We examine two different methods: a Convolutional Neural Network (CNN) approach and a Transfer Learning (TL) approach and we compare their performance with and without augmentation. The results confirm an overall enhancement using augmentation methods with deep learning classification methods (especially transfer learning) when evaluated on the two datasets.**

***Keywords*—Generative Adversarial Networks (GAN); Convolutional Neural Network (CNN); deep learning; breast cancer; Transfer Learning (TL); data augmentation; ultrasound (US) imaging; cancer diagnosis***

## I. Introduction

Medical imaging is a worthy tool to diagnose the presence of several diseases and the analyize of the experimental results [1]. Biomedical imaging is part of the foundations of overall cancer care. Breast cancer is well-known and widespread through women world-wide and it causes mortality rates. It is anticipated that more than eight percent of women will acquire breast cancer during their lifetime [2]. Digital Mammography (DM) is the most generally used and practical technique for breast cancer diagnosis [3]. Early detection is the most important factor in decreasing the costs of cancer management and mortality. DM imaging has some weaknesses in dense breasts where tumors can be hidden by surrounding tissue (where the dense tissue has a similar attenuation contrasted to the tumor). In practice, ultrasound (US) imaging is the best alternative to DM, which is applied as a complementary approach for breast cancer classification and detection due to its sensitivity, safety and versatility [4]. However, the weakness of US imaging is that it is hand-dependent which relies more on radiologists. Explaining US images needs specialist radiologists due to its difficulty and appearance of speckle noise. Therefore, Computer Aided Diagnosis (CAD) can help radiologists in the US-based classification and detection of breast cancer, reducing the influence of the hand-dependence of US imaging.

Some researches have studied the effect of CAD diagnostics [5], [6] and noted that CAD is a robust tool to enhance diagnostic specificity and sensitivity. Breast US research has a shortage of benchmark dataset which results in limiting the advancement of recent algorithms. Therefore, breast US images quality is extremely dependent on the acquisition process and there is a large variability between various US systems that affects the outputs achieved by algorithms. The output is also influenced by the size, location, and appearance of the tumor or micro-calcification.

Training a deep model on insufficient data regularly results in over-fitting because a model of high capacity is capable of "memorizing" the training set. Multiple methods have been presented to mitigate this problem, but none performed effectively as to be used exclusively. These techniques can be split into two large categories: (1) regularization techniques, pointing to limit the model's capacity (e.g., dropout and parameter norm penalty) and (2) data augmentation techniques, aiming to increase the size of the dataset [7]. In practice, most models improve from these two techniques. We concentrate on these two categories. GANs [8] are a family of unsupervised neural networks most generally utilized for image generation. Data augmentation has confirmed to be very efficient and is adopted universally in the field of deep learning [9], [10]. It is in fact so effective that it is being used even in tasks that include massive data [11]. The most common forms of augmentation include flipping, scaling, translating, rotating, blurring and sharpening. The goal of such transformations is to obtain a new image that contains the same semantic information as the original.

While augmentation most certainly helps neural networks

learn and generalize more effectively, it also has its drawbacks. In most cases, augmentation techniques are limited to minor changes on an image, as more "heavy" augmentations might damage the image's semantic content. Furthermore, the forms of augmentation one can use differ from problem to problem, making their application ad-hoc and empirical. For instance, medical images have to be mildly augmented as they follow strict standards (i.e., they are centered, their orientation and intensity vary little from image to image and many times they are laterally/horizontally asymmetric) [12]. Finally, augmentation techniques are applied to one image at a time and thus are unable to gather any information from the rest of the dataset.

### A. Problem Statement and Motivation

Although there are a lot of scientific researches in the process of classification and detection of cancer tumors using different types of modalities, breast US imaging has rare researches due to the shortage of public benchmark datasets. We utilize Data Augmentation Generative Adversarial Networks (DAGANs) [13] to make our dataset (BUSI) and dataset (B) larger. We particularly chose to use breast US imaging because US scan is safe for human body while DM and other screening technology may not achieve the same standard of safety as US imaging. Furthermore, We are proposing deep learning approaches for breast US imaging classification using state-of-the-art algorithms to improve the accuracy results using deep learning approaches proved to achieve promising results.

### B. Paper Contribution

- Due to the scarce number of datasets of ultrasound images for breast cancer, we believe that our dataset collection and data augmentation are an important contribution that can be a great seed for related studies. We plan to make our dataset publicly available for other researchers.
- We propose a novel augmentation technique that overcomes the above-mentioned limitations and is capable of augmenting any given dataset with realistic, high-quality images generated from scratch using DAGAN.
- We used two datasets which are our BUSI dataset and dataset B [14]. And we ran state-of-the-art deep learning models; CNN and TL as classification algorithms and they produced promising results.
- Finally, We merge the two datasets to overcome the limitation of the size of the dataset and compare the new results (the merged dataset) with the previous results (the two separate datasets). In addition, we will enlarge our datasets by combining them with traditional augmentation and DAGAN data to enhance the final results.

The remainder of this paper is divided as follows: Section II clarifies some related work in these fields. Subsequently, Section III illustrates the two Breast US datasets. Section IV discusses our methodology. Section V contains the results and discussion. And finally, Section VI has a conclusion and future work.

## II. Related Work

In this section, related work for breast US image classification and data augmentation in medical images are reviewed. Furthermore, a brief introduction about deep learning for breast imaging is discussed.

### A. Breast US Image Classification

This section explains in brief three state-of-the-art approaches for tumor classification in breast US imaging.

*1) Convolutional Neural Networks (CNN):* Huynh et al. [15] assessed the performance of utilizing transferred features from pretrained CNNs [16] in classifying cancer in breast US images, and to examine this method of transfer learning with preceding methods including human-designed features. A breast US dataset composed of 1125 samples and 2392 Regions of Interest (ROIs) was utilized. Every ROI was annotated as malignant or benign. Features were extracted from each ROI using pre-trained CNNs and used to train Support Vector Machine (SVM) classifiers in the tasks of distinguishing benign vs malignant tumors. For a baseline comparison, classifiers were also trained on prior analytically-extracted tumor features. They conducted five-fold cross-validation with the area under the receiver operating characteristic curve (AUROC) as the performance metric. Classifiers trained on CNN-extracted features were comparable to classifiers trained on human-designed features. In the task of malignant versus benign, the SVM trained on both CNN-extracted features and human-designed features achieved an AUC of 90%. In the task of determining benign vs malignant, the SVM trained on human-designed features achieved an AUC of 85%, compared to the AUC of 85% achieved by the SVM trained on CNN-extracted features. The authors obtained great results using transfer learning to characterize ultrasound breast cancer images. This method allows them to instantly classify a little dataset of lesions in a computationally reasonable fashion without any hand-operated input. Current deep learning approaches are dependent on huge datasets and large computational resources, which are frequently difficult to access for clinical applications. It is important to highlight that, the dataset of this study [15] is not publicly available neither by request.

*2) Stacked Deep Polynomial Network (S-DPN):* Jun Shi et al. [17] proposed Deep Polynomial Network (DPN) [18] algorithm not just presents better performance on a massive dataset, but also has the possibility to learn strong characteristic representations from a comparatively little dataset. In their study, a S-DPN algorithm is suggested to further enhance the representation performance of the primary DPN, and S-DPN is then used to the task of texture feature learning for US classification of tumor with a little dataset. The task of tumor classification is achieved on two datasets, namely the prostate US elastography dataset and breast B-mode US dataset. On these two cases, results of the experiment confirm that S-DPN achieves the best classification performance with accuracies of 92.40% on breast US dataset and 90.28% on prostate US datasets. It is important to highlight that, the dataset of this study [17] is not publicly available.

*3) Shearlet-based Texture Feature Extraction:* Zhou et al. [19] augmented the classification accuracy of the US computer-aided diagnosis (CAD) for the detection of breast tumor

based on texture feature, they also offered to use Shearlet transform to achieve texture feature descriptors. Shearlet transform produces a scattered representation of high-dimensional data with especially higher directional sensitivity at different scales. Hence, texture feature descriptors of shearlet-based can strongly explain breast tumors. In order to accurately evaluate the achievement of Shearlet-based features, curvelet, contourlet, and wavelet-based texture feature descriptors are also obtained for comparison. All these features were then fed to two different classifiers, AdaBoost and support vector machine (SVM), to estimate the consistency. The results of the experiment of breast tumor classification presented that the classification accuracy, specificity, sensitivity, negative predictive value, positive predictive value and Matthew's correlation coefficient of shearlet-based method were 91.0%, 92.5%, 90.0%, 90.3%, 92.6%, 0.822% by SVM, and 90.0%, 90.0%, 90.0%, 89.9%, 90.1%, 0.803% by AdaBoost, respectively. Most of the results of the Shearlet-based significantly exceeded those of other approach based results under both classifiers. They suggested a new texture feature extraction approach based on Shearlet transform for describing breast tumor in US image. The comparative experiment results showed that the Shearlet-based texture feature can more efficiently identify breast tumors in US image than other features extracted from curvelet, contourlet, wavelet and Gray-Level Co-Occurrence Matrix (GLCM) approaches. It is important to highlight that, the dataset of this study [19] is not publicly available.

### B. Data Augmentation

GANs have been successfully used for data augmentation. Wang et al. [20] and Antoniou et al. [13], for example, use custom GAN architectures in a low-data setting to achieve consistently better results than traditionally augmented classifiers, while Perez et al. [21] devise a novel pipeline called Neural Augmentation which, through style transfer techniques, aims at generating images of different styles, performing equally as good as traditional augmentation schemes in a subsequent classification task. Additionally, Neff [22] proposes a generative model which learns to produce pairs of images and their respective segmentation masks in order to assist a UNet segmentation model, proving that in simpler datasets networks trained with a mix of synthetic and real images stay competitive with networks trained on strictly real data using usual data augmentation.

One field in which data augmentation is especially important is that of medical imaging, where the lack of available public data is a ubiquitous problem since access to individual medical records is heavily protected by legislation and appropriate consent must be given. In most cases, this process is hindered by bureaucracy and/or high costs, while the resulting collection is greatly imbalanced towards normal subjects. Several authors employ Machine Learning techniques to learn directly from the available data and surpass the state-of-the-art in problems as diverse as generating benchmark data, cross-modality synthesis, super-resolution or image normalization [23].

The medical field has only recently started adopting GAN-based methodologies for synthesizing images [24]. In particular, Bentaieb et al. [25] and Shaban et al. [26] proposed GAN-based style transfer approaches to stain normalization in histopathology images, with quite interesting results in various datasets. For tackling segmentation tasks, various authors have proposed custom GAN architectures and pipelines which are adversarially trained to produce proper segmentation masks from a given medical image dataset [27]–[29]. Regarding image translation between modes, the authors of [30] synthesize T2-weighted brain MRI images from T1-weighted ones, and vice versa, using a Conditional GAN model. Finally, many authors, such as [31] and [32], have attempted to generate counterfeit medical images in order to increase the size of the training set of different deep learning models, a task more closely related to the one examined in this study.

Supplementary to all of the above efforts, our approach aims to exploit the superior performance of GANs for the benefit of medical image classification. We explore the impact of GAN-assisted data augmentation on the diagnosis of breast cancer through US scans.

### C. Deep Learning for Breast Imaging

In general, the state-of-the-art classification methods are not robust, specifically the image processing based methods, relying on special assumptions and rule-based methods. Without necessitating such a powerful hypothesis, deep learning approaches have shown an improved accuracy in object classification and detection, which proposed that could also improve the state-of-the-art of tumor classification in breast ultrasound. Deep learning in medical imaging is usually represented by convolutional networks. GANs [8] are a family of unsupervised neural networks most usually used for image production. Each GAN is formed of two networks: a generator and a discriminator, playing against each other in a two-player game. These models have proven to be capable of creating realistic images and will serve as an assisting basis for this study. DAGAN is also used to make the dataset larger. Based on how we can train them, they can be frequently categorized into the following categories:

1) **CNNs approach.** This method trains the CNNs with images for training and testing [33], [34]. However, feeding every image to the network is time-consuming [35].
2) **Transfer learning approach.** Another approach that has been extensively used recently in biomedical research is the transfer learning technique [15], [36]. This method uses a pretrained model from natural images to overcome the lack of data in medical imaging study.
3) **Generative Adversarial Networks.** This method allows us to generate new images from our dataset. GAN [8] is a strong and new approach in image synthesizing.

In breast imaging, the majority of the current publications are focusing on using CNNs for MG. Dhungel et al. [37] have performed masses segmentation using deep learning; Mordang et al. [38] introduced the use of CNNs in microcalcification detection; and lately, Ahn et al. [39] suggested the use of CNNs in breast density evaluation. In breast US imaging, Huynh et al. [15] suggested the use of a transfer learning approach for breast US images classification. Yap et al. [14] proposed to use deep learning approaches for classification

of breast US tumor. As of the date of this publication, this is the only work the authors have found that handles breast ultrasound but it does not enhance the accuracy in tumor classification. Most of the aforementioned work focused on lesion detection. Furthermore, publications utilizing data augmentation with GAN are rare. In medical images, Frid-Adar et al. [31] proposed the use of DAGAN to enhance CNN performance in liver lesion classification. We are, in our consideration, the first to use DAGAN with breast US images. In this paper, we propose to use deep learning approaches for breast US tumors classification. To show the benefits of deep learning approaches, we compare the performances among all the deep learning approaches which are used in this paper for tumor classification. Furthermore, DAGAN and traditional augmentation are used to make the dataset larger and enhance the performance of our classification approaches.

## III. Datasets

In general, to develop a healthcare system using deep learning, a dataset should be available. This study uses two different datasets of breast US images. Our dataset BUSI was collected and obtained from US systems with different specifications and at different times. The Dataset B [14] was requested from its owners. Examples of both datasets are shown in Fig. 1.

Dataset BUSI collected at baseline includes ultrasound breast images among women in ages between 25 to 75 years old. The number of patients is 600 female patients. It was collected in 2018 from Baheya Hospital for Early Detection and Treatment of Women's Cancer, Cairo (Egypt) with LOGIQ E9 ultrasound system and LOGIQ E9 Agile ultrasound. The data is categorized into three classes, which are normal, benign, and malignant. The dataset consists of 780 images from different women with an average image size of 500 x 500 pixels. Within the 780 tumor images, 133 were normal images without cancerous masses, 437 were images with cancerous masses and 210 were images with benign masses. Our dataset BUSI is available online[1] for studies.

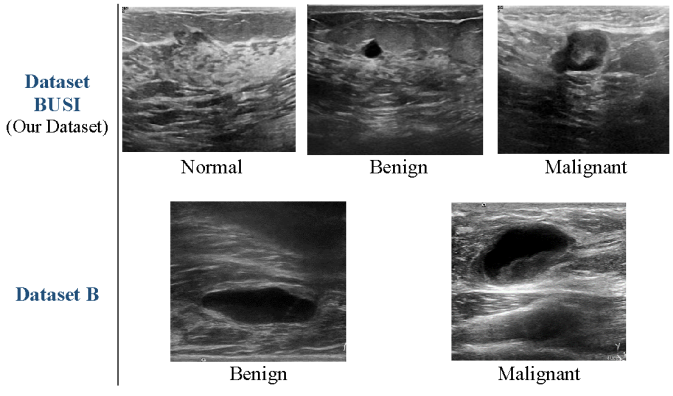


Fig. 1. Samples of breast US images from both datasets where the first row contains images from dataset BUSI and the second row contains images from dataset B.

The other dataset is referred to as Dataset B [14]. It was collected in 2012 from the UDIAT Diagnostic Center of the Parc Tauli Corporation, Sabadell (Spain). It has 163 images from different females. The average image size of the dataset is 760 x 570 pixels. The number of images in the dataset is 163 images where 53 images were with malignant masses and 110 images were with benign tumors. It was created for lesion detection not for classification while our study uses it for lesion classification.

## IV. Methodology

Our methodology is divided into two parts. In the first part, we discuss data augmentation using GAN and traditional augmentation. While the second part discusses classification techniques which are performed by using deep learning approaches Convolutional Neural Network (CNN) and a Transfer Learning (TL) on BUSI dataset, dataset B and merged datasets (BUSI+B). The whole model architecture is shown in Fig. 2. It is important to highlight that the classification algorithms were performed on four forms of data samples as follows: (1) without augmentation which means the real images(the blue dash line and arrows). (2) with traditional augmentation. (3) using DAGAN. (4) using traditional augmentation and DAGAN ( the orange box and arrows).

### A. Data Augmentation Generative Adversarial Networks (DAGAN)

The second goal of this study was to produce realistic images for each of the classes on-demand while the first goal is to enhance the classification accuracy using deep learning approaches. Each GAN is composed of two networks: a generator and a discriminator, playing against each other in a two-player game. These models have proven to be capable of creating realistic images. To achieve this, a framework was performed where a single GAN was trained on each of the classes. A GAN architecture of sufficient capacity to understand and model the underlying distributions of each of the classes had to be selected. A GAN that satisfies the above goal should, after training, be able to produce realistic images of the class it was trained upon.

Furthermore, GAN [8] is formed of two networks which are the generator and the discriminator. The generator accepts a noise vector as input and produces fake data, which are then fed, along with real ones, to the discriminator, whose goal is to distinguish which distribution the samples were produced from. Conversely, the generator's goal is to learn the real distribution without witnessing it, in order to make its output indistinguishable from real samples. Both networks are trained simultaneously and adversarially until an equilibrium is reached. In order to combat instability issues during training, the Earth Mover's or Wasserstein distance was used, partially because it leads to convergence for a much broader set of distributions, but mostly because its value is directly correlated to the quality of the generated data [40]. The discriminator was initially achieved by clipping its weights by an arbitrary value Wasserstein GAN (WGAN) [40]. It was later shown that this technique led to sub-optimal behavior, which could be ameliorated with the inclusion of a gradient penalty term to the discriminator's loss function calculated on a random interpolation point between the real and the fake samples [41]. The resulting architecture WGAN gradient penalty (WGAN-GP) [41] is the one utilized in our study.

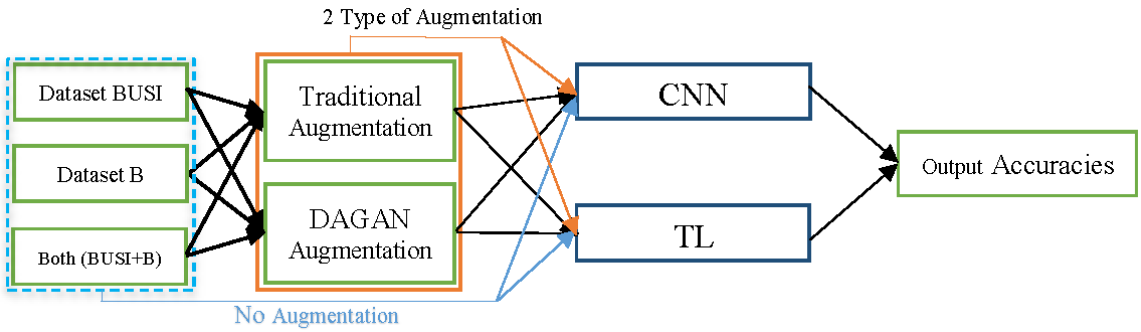


Fig. 2. The proposed methods for breast US image classification and data augmentation techniques.

*1) Generator:* An architecture with 11 layers was selected as the generator of the network. The architecture is depicted in Fig. 3A. The generator input is a vector of 128 random values in the range of (0,1). It is sampled from a uniformed distribution. A Fully Connected (FC) layer followed the input layer. The subsequent layers are regular 2D convolutions (Conv) and 2D transposed convolutions (Conv trans up), sometimes referred to as "deconvolution" layers. A 5x5 sized kernel and "same" padding were selected for both types of layers, while a stride of 2 was selected for the transposed convolutions. This performs in the doubling of the spatial dimensions of its input. A "Leaky ReLU" function activated all layers apart from the last layer. The final layer has a tangent hyperbolic (tanh) activation function because its output needs to be bound in order to be able to output an image. A tanh function was preferred over a sigmoid function because it is centered around 0, which helps during training [42]. Finally, after five alternations of convolution and transposed convolution layers (each of which doubles the size of its input), an image with a resolution of (192x160) and 1 channel is produced.


Fig. 3. DAGAN architecture: A) Generator structure and B) Discriminator structure

*2) Discriminator:* The discriminator is a usual CNN architecture intended towards binary classification. The one used in the present study consists of 11 layers can be seen in Fig. 3B. The input to the discriminator is a single-channel 192x160 image. This image is then passed five times through alternating layers of convolutions with a stride of 1 and 2 respectively; the latter is used for sub-sampling as there are no pooling layers present in the architecture. The last two layers are FC ones. All layers in the network are activated by a "Leaky ReLU" , besides the last one which has no activation function.

### B. Traditional Augmentation Techniques

Due to the nature of our datasets, we could only implement a limited range of visual transformations. In particular, we applied a horizontal flip, brightness, scaling and zooming. The number of augmented images that were obtained from traditional augmentation would increase by a factor of 2 for each augmented method.

### C. Convolutional Neural Network

Based on the Deep learning definition, it is a representation learning approach [43] that will automatically detect features satisfying a special task from the data. The feature extractors are task-specific, in that they are not fixed to a set of specific rules every time [44]. Each network contains multiple layers that lead to hierarchical features used in the learning process [43], [45].

CNNs [46] are a valuable technique in image analysis, particularly in recognition, detection or classification of faces [47], text [45], biological images [48] and human bodies [49]. For these reasons, we study the performance of deep learning in breast US tumor classification.

CNNs consist of convolutional layers and pooling layers [46], where the role of the former is to extract local features from a set of learnable filters and the role of the latter is to merge neighboring patterns, reducing the spatial size of the previous representation and adding spatial invariance to translation [43]. CNNs are hierarchical neural networks and their accuracy is based on the design of the layers and training models [50].

Some common CNNs are available which are AlexNet [16], LeNet [45] and GoogleNet [51]. We studied the use of two types of deep learning models for breast classification: AlexNet [16] and a transfer learning approach using Convolutional Networks [52].

*1) CNN-AlexNet:* As the ultrasound breast images in the datasets are gray-scale and the size of the breast tumor or micro-calcification is relatively small, AlexNet [16] was chosen as a suitable architecture to solve the classification problem of multi-classes. The training and validation images are input of the model containing all classes in the datasets. We split all datasets to 70%,15%, and 15% for training, validation, and testing, respectively. The AlexNet architecture is simple and was primarily built for digit classification [45]. Breast tumors include related gradients that can be presented through CNNs. The overall architecture is shown in Fig. 4, with the inputs consisting of images of breast tumors and normal tissue. The inputs are fed into the first convolution layer and max-pooling layer, which is repeated once and finalized with two fully connected layers. The final number of outputs are 2 neurons or 3 neurons, which are the activations generated for the two or three classes: (benign and malignant) or (normal, benign and malignant), respectively. The final part of the CNN is the output of class probabilities to measure how close the final fully connected parameters are with respect to the labels of the training and validation data. The loss was calculated using multinomial logistic loss with a softmax classifier. The output of our network is a prediction of whether the image is a tumor or healthy breast tissue. It is formed by two fully connected layers with the softmax function defined as

$$f_j(z) = \frac{e^{z_j}}{\sum_k e^{z_k}} \tag{1}$$

where $f_j$ is the *j*-th element of the vector of class scores $f$ and $z$ is a vector of random real-valued scores that are flattened to a vector of values between zero and one that sum to one. The loss function is defined so that having good predictions during training is equivalent to having a small loss. A Rectified Linear Unit (ReLU) layer is included at the first fully-connected layer. This element-wise operation is calculated and defined as

$$f(x) = max(0, x) \tag{2}$$

where the function $f$ thresholds the activations at zero.

*2) Transfer Learning:* Transfer Learning (TL) [53], [54] is a method where a CNN is trained to learn features for a broad domain after which the classification function is changed to optimize the network to learn features of a more specific domain. Under this setting, the features and the network parameters are transferred from the broad domain to the specific one. Furthermore, Transfer Learning (TL) is a method that provides a system to apply the knowledge learned of prior tasks to a new task domain that is somehow related to the prior domain. Our proposed transfer learning approach is based on VGG16 [55], ResNet [56], Inception [57], and NASNet [58]. These networks were primarily utilized for the classification of more than one thousand various objects of classes on the ImageNet dataset [16]. The default image sizes for TL models are shown in Table I.

TABLE I. THE DEFAULT INPUT SIZE FOR TRANSFER LEARNING MODELS

| Models | Input Shape |
|---|---|
| VGG16 | 224x224 |
| ResNet | 224x224 |
| Inception | 299x299 |
| NASNetLarge | 331x331 |

### D. Implementation

*1) Preprocessing:* In this subsection, we focus on the preparation of the datasets and image augmentation. Additional preprocessing steps were taken to facilitate model training, such as resizing them to 192x160 with Lanczos interpolation. In addition, we randomly divided the dataset into training, validation and test sets, keeping intact the sequence of each image so that every image appears in only one of the aforementioned sets. We should note here that in our initial experiments we randomly shuffled and split all images without preserving each image sequences; this allowed the models to identify key features in each subject's morphology and achieve a perfect score on the test set (i.e., for each test set image, the model had been trained on another from the same image). Because of this, the study of the models' generalization on new, unseen patients, which is a necessary requirement for all medical applications, became infeasible. We train both datasets (BUSI and B) using DAGAN model. Trained models are saved and reused in generating new images. DAGAN runs in 700 epochs for each class (normal, benign and malignant). Samples of real image for datasets and augmented images are shown in Fig. 5. We generate 5000 images for each class using DAGAN model. All the images are added to our datasets.

*2) Classification Methods:* The proposed CNN approach in this paper is AlexNet model [16]. The breast US images are in grayscale. The datasets were split into 70%, 15%, and 15% for training, validation, and testing, respectively. The validation set (15%) is used for hyper-parameter tuning and early stopping. The network is trained by using Adam optimizer - with a learning rate of 0.0001. It uses 60 epochs (early stopping) with 0.30 of dropout rate. We used a stride of one and two pixels in max-pooling. To obtain the best performance for the state-of-the-art classification methods on the datasets, we use regularization techniques such as normalization and dropout.

For transfer learning, we used four pretrained models which are VGG16 [55], ResNet [56], Inception [57], and NASNet [58]. An Adam optimizer is used with learning rate 0.001. The number of epochs was 10 epochs. The output layer of TL models is altered and we train our data in it. The softmax activation function is utilized in TL experiments.

In order to measure the effectiveness of the proposed methodology, the following experiment was devised: Firstly, a Deep neural network architecture was selected, which is capable of achieving satisfactory performance on classifying the three classes (i.e. normal, benign and malignant). Secondly, DAGAN and traditional augmentation are used to enhance the performance of classification algorithms by generating more data-samples.

Our methodology is summarized in the following points:

1) First, we have two datasets BUSI and dataset B and a third one that was created by merging the two datasets (BUSI+B).
2) We perform two types of data augmentation to generate more data samples, the first type is traditional augmentation and the second type is DAGAN.

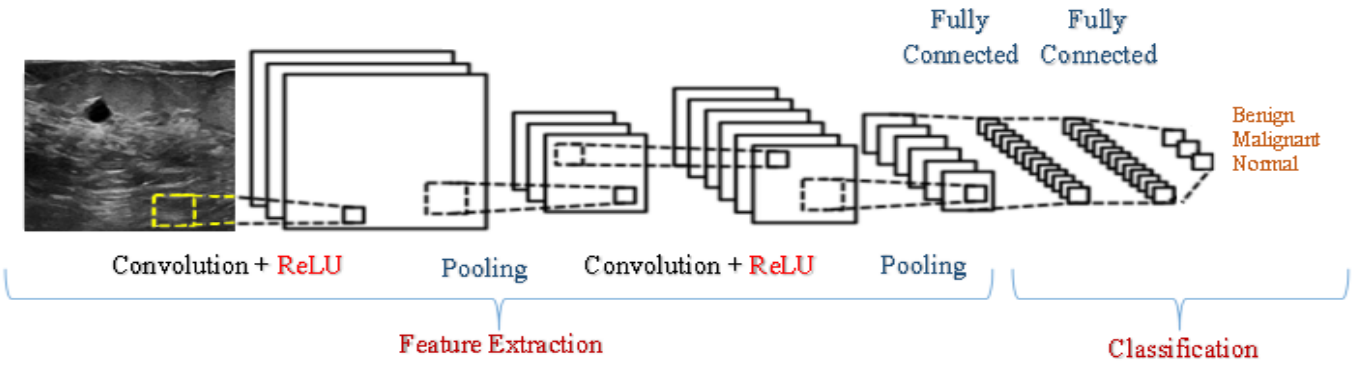


Fig. 4. CNN architecture of AlexNet Model.

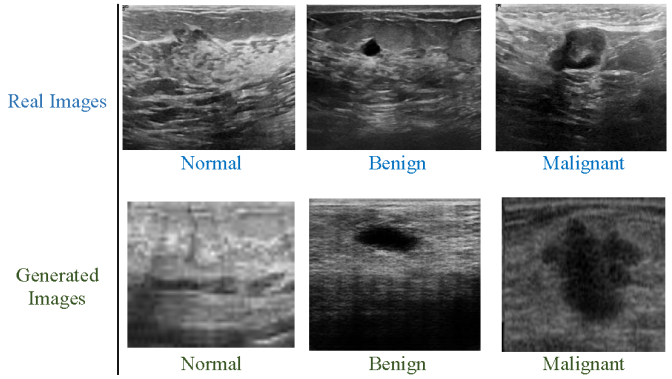


Fig. 5. Samples of real images from dataset BUSI are in the first row and augmented images using DAGAN are in the second row.

3) Two deep learning classification approaches were used, CNN (AlexNet) and TL (VGG16, ResNet, Inception, and NASNet).
4) It is important to highlight that in our experiments we perform the classification algorithms on four forms of data samples as follows: (1) without augmentation which means the real images. (2) with traditional augmentation. (3) using DAGAN. (4) using traditional augmentation and DAGAN. As a result of this, the total number of 60 classification codes have been implemented (see Table II and Fig. 6)

### E. Implementation Environment

Our classification experiments are performed on Windows 10 operating system using Keras API library[2] version 2.0.1 (on top of TensorFlow[3]) using Python (version 3.5). In this study, training and classification are performed on Intel (R) Core (TM) i73630QM CPU @2.40MGz and GPU NVIDIA Quadro K2000M With 8GB of shared GPU memory, and 16 GB RAM.

Furthermore, DAGANs are performed in a powerful server which uses Ubuntu 18.04 operating system using as mentioned above, Tensorflow, Keras, and Python. The server specification is Intel Xeon(R) CPU E5-26200 @2.00GHz×12, llvmpipe (LLVM 7.0, 256bits), and 50 GB RAM.

[2]https://keras.io/
[3]https://www.tensorflow.org

## V. Results and Discussion

There are many parameters that affect the results in deep learning when used in medical images such as the type of algorithms, hyperparameters, and size of the dataset. We considered all of these parameters in our experiments.

Dataset BUSI was obtained from a modern US system, which offers new challenges for the current techniques in tumor classification. These US systems obtain high-resolution images that may cover other structures such as air in the lungs, ribs or pectoral muscle, making the tumor classification more difficult. Dataset B was collected from an older US system. Images are usually of a lower resolution. However, these differences did not affect our experiments.

The performance accuracies of our experiments are shown in Table II. Regarding our experiments, we found the following:

1) In all of our experiments, we found that increasing the number of data samples using data augmentation and datasets merging, significantly improve the classification accuracies. This is obvious in the results Table II and Fig. 6. Note that the classification results obtained from DAGAN outperform the traditional argumentation. While the results were the best when we combine DAGAN and traditional argumentation.
2) When we performed the experiments on datasets without data augmentation, they produce low accuracies (even if we combined the two datasets (BUSI+B)). This is due to the shortage of data.
3) We figured out that traditional augmentation is not very effective in our work due to the nature of medical images. In addition, medical images are not like natural images that are used in object classification. There are limited traditional augmentation techniques that can be used in medical images.

These results showed that the supervised deep learning methods were data-driven and the performance increased with more training dataset. We can confirm that the transfer learning approach achieved the best accuracy when trained with data augmentation through the use of DAGANs and traditional

TABLE II. COMPARISON OF THE ACCURACY OF DIFFERENT METHODS WHEN TESTING ON SINGLE AND COMBINED DATASET. THE BEST RESULTS IS INDICATED IN BOLD.

| Dataset | Method | Sub-Method | Without Augmentation | Traditional Augmentation | DAGAN Augmentation | Both Traditional and DAGAN augmentation |
|---|---|---|---|---|---|---|
| Dataset BUSI | CNN-AlexNet | | 58% | 62% | 73% | 78% |
| | TL | VGG16 | 70% | 74% | 84% | 88% |
| | | Inception | 68% | 73% | 82% | 85% |
| | | ResNet | 79% | 82% | 89% | 93% |
| | | NASNet | 83% | 85% | 91% | **94%** |
| Dataset B | CNN-AlexNet | | over-fitting | 56% | 75% | 80% |
| | TL | VGG16 | 68% | 72% | 80% | 82% |
| | | Inception | 65% | 70% | 77% | 80% |
| | | ResNet | 75% | 79% | 86% | 90% |
| | | NASNet | 79% | 82% | 90% | **92%** |
| Datasets (BUSI+B) | CNN-AlexNet | | 60% | 65% | 82% | 84% |
| | TL | VGG16 | 72% | 75% | 86% | 88% |
| | | Inception | 70% | 73% | 84% | 87% |
| | | ResNet | 76% | 79% | 88% | 92% |
| | | NASNet | 84% | 88% | 96% | **99%** |

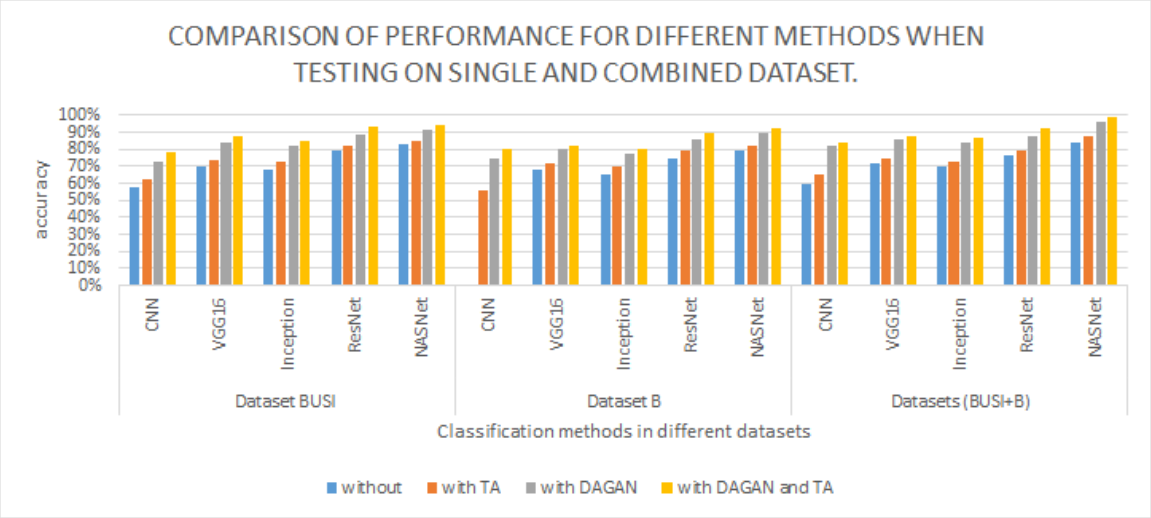


Fig. 6. The chart illustrates the performance accuracies in all the performed methods with three datasets in (without: no data augmentation, with TA: with traditional augmentation, with DAGAN: using generated images, with DAGAN and TA: used traditional augmentation and generated images).

augmentation in combined dataset (BUSI+B). The final result is 99% (when training on TL NASNet pretrained model).

The use of a powerful generative model for producing images (e.g., DAGAN) has many advantages over traditional augmentation schemes. The most important advantage is the quality of the produced images and the capability of generalizing beyond the limits of the original dataset to produce new patterns. The proposed technique is especially useful in low-variance datasets where the images follow a very strict format. We would like to point out that other studies reached 92% accuracy in classification methods using their own datasets while we reached 99% using our datasets.

On the other hand, there are some limitations in our work which are:

- The training process is time consuming and requires high computer resources.
- There is not a sufficient number of real images that have been collected to avoid the classification errors in the augmented images.
- We can not synthesize high-resolution images using DAGAN.

## VI. CONCLUSIONS AND FUTURE WORKS

This paper investigated the use of two deep learning classification approaches (particularly CNN (AlexNet) and Transfer Learning approaches). Two datasets were used which are our Dataset BUSI and Dataset B. Furthermore, we combined them obtaining a third one which is dataset (BUSI+B). We used a novel methodology for data augmentation with the use of GAN. It involves training a GAN for each of the classes of the original datasets and then using it to produce a number of synthetic images. All models were trained on breast US images datasets to classify cancerous and non-cancerous images.

To study the impact of this augmentation strategy for classification methods, four experiments were conducted. Firstly, CNN and transfer learning models were trained on all datasets

on a form of baseline. Secondly, the same models were trained with traditional and thirdly by the proposed GAN augmentation techniques. Fourthly, by both forms of augmentation(traditional and DAGAN).

The performances were evaluated on the three datasets (BUSI, B, and BUSI+B). Amongst the various methodologies presented in this paper, the transfer learning NASNet achieved the best results (99%) in Dataset (BUSI+B) when it is used with DAGAN and traditional augmentation. Deep learning methods are adaptable to the specific characteristics of any dataset since these are machine-learning based and particular models are constructed for each dataset. Experiments confirm that augmentation through GANs outperforms traditional augmentation methods when used with CNN and transfer learning.

Finally, the models trained with the proposed methods using GAN augmentation methodology outperform the ones with a traditional one by a large margin. In fact, because of the nature of the images, the traditional techniques gave no enhancement over the baseline experiments. The final experiments, which combined both forms of augmentation exceeded the rest, pointing that while traditional augmentation could not function on its own, it performs well when it is combined with GAN augmentation.

In the future, we believe that deep learning approaches could be adjusted to other medical imaging techniques such as 3D ultrasound or other modalities. Mass classification is the initial step of a CAD system. Hence, in our future work we plan to do breast ultrasound lesion detection and segmentation, and evaluate the performance of the complete CAD system. Because of the improved results of our experiments using DAGAN, multiple future research areas could be spawned. We are planning to experiment with different structures for further developments on the data quality, either within the WGAN-GP by utilizing a more robust discriminator, or by using a newer, more modern framework that leads to enhanced experimental performance, such as the Progressive Growing GANs [59] or the Auxiliary Classifier GANs [60].

## REFERENCES

[1] E. Hall, "Radiobiology for the radiologist, radiation research, vol. 116, no. 1, 1988."

[2] H.-D. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: A survey," *Pattern recognition*, vol. 43, no. 1, pp. 299–317, 2010.

[3] O. Akin, S. B. Brennan, D. D. Dershaw, M. S. Ginsberg, M. J. Gollub, H. Schöder, D. M. Panicek, and H. Hricak, "Advances in oncologic imaging: update on 5 common cancers," *CA: a cancer journal for clinicians*, vol. 62, no. 6, pp. 364–393, 2012.

[4] A. T. Stavros, D. Thickman, C. L. Rapp, M. A. Dennis, S. H. Parker, and G. A. Sisney, "Solid breast nodules: use of sonography to distinguish between benign and malignant lesions." *Radiology*, vol. 196, no. 1, pp. 123–134, 1995.

[5] M. H. Yap, E. Edirisinghe, and H. Bez, "Processed images in human perception: A case study in ultrasound breast imaging," *European journal of radiology*, vol. 73, no. 3, pp. 682–687, 2010.

[6] K. Drukker, N. P. Gruszauskas, C. A. Sennett, and M. L. Giger, "Breast us computer-aided diagnosis workstation: performance with a large clinical diagnostic population," *Radiology*, vol. 248, no. 2, pp. 392–397, 2008.

[7] J. Kukačka, V. Golkov, and D. Cremers, "Regularization for deep learning: A taxonomy," *arXiv preprint arXiv:1710.10686*, 2017.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[9] D. C. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Deep, big, simple neural nets for handwritten digit recognition," *Neural computation*, vol. 22, no. 12, pp. 3207–3220, 2010.

[10] C. N. Vasconcelos and B. N. Vasconcelos, "Increasing deep learning melanoma classification by classical and expert knowledge based image transforms," *CoRR, abs/1702.07025*, vol. 1, 2017.

[11] R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun, "Deep image: Scaling up image recognition," *arXiv preprint arXiv:1501.02876*, 2015.

[12] Z. Hussain, F. Gimenez, D. Yi, and D. Rubin, "Differential data augmentation techniques for medical imaging classification tasks," in *AMIA Annual Symposium Proceedings*, vol. 2017. American Medical Informatics Association, 2017, p. 979.

[13] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *arXiv preprint arXiv:1711.04340*, 2017.

[14] M. H. Yap, G. Pons, J. Martí, S. Ganau, M. Sentís, R. Zwiggelaar, A. K. Davison, and R. Martí, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE journal of biomedical and health informatics*, vol. 22, no. 4, pp. 1218–1226, 2018.

[15] B. Huynh, K. Drukker, and M. Giger, "Mo-de-207b-06: Computer-aided diagnosis of breast ultrasound images using transfer learning from deep convolutional neural networks," *Medical physics*, vol. 43, no. 6Part30, pp. 3705–3705, 2016.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[17] J. Shi, S. Zhou, X. Liu, Q. Zhang, M. Lu, and T. Wang, "Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset," *Neurocomputing*, vol. 194, pp. 87–94, 2016.

[18] R. Livni, S. Shalev-Shwartz, and O. Shamir, "An algorithm for training polynomial networks," *arXiv preprint arXiv:1304.7045*, 2013.

[19] S. Zhou, J. Shi, J. Zhu, Y. Cai, and R. Wang, "Shearlet-based texture feature extraction for classification of breast tumor in ultrasound image," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 688–696, 2013.

[20] Y.-X. Wang, R. Girshick, M. Hebert, and B. Hariharan, "Low-shot learning from imaginary data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7278–7286.

[21] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[22] T. Neff, C. Payer, D. Stern, and M. Urschler, "Generative adversarial network based synthesis for supervised medical image segmentation," in *Proc. OAGM and ARW Joint Workshop*, 2017.

[23] A. F. Frangi, S. A. Tsaftaris, and J. L. Prince, "Simulation and synthesis in medical imaging," *IEEE transactions on medical imaging*, vol. 37, no. 3, p. 673, 2018.

[24] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *arXiv preprint arXiv:1809.07294*, 2018.

[25] A. BenTaieb and G. Hamarneh, "Adversarial stain transfer for histopathology image analysis," *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 792–802, 2018.

[26] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay, "Gans for medical image analysis," *arXiv preprint arXiv:1809.06222*, 2018.

[27] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *International Workshop on Simulation and Synthesis in Medical Imaging*. Springer, 2018, pp. 1–11.

[28] W. Dai, X. Liang, H. Zhang, E. Xing, and J. Doyle, "Structure correcting adversarial network for chest x-rays organ segmentation," Sep. 27 2018, uS Patent App. 15/925,998.

[29] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang, "Segan: Adversar-

ial network with multi-scale l 1 loss for medical image segmentation," *Neuroinformatics*, vol. 16, no. 3-4, pp. 383–392, 2018.

[30] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, "Image synthesis in multi-contrast mri with conditional generative adversarial networks," *IEEE transactions on medical imaging*, 2019.

[31] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[32] P. Costa, A. Galdran, M. I. Meyer, M. Niemeijer, M. Abràmoff, A. M. Mendonça, and A. Campilho, "End-to-end adversarial retinal image synthesis," *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 781–791, 2018.

[33] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in neural information processing systems*, 2012, pp. 2843–2851.

[34] T. Kooi, G. Litjens, B. Van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer, "Large scale deep learning for computer aided detection of mammographic lesions," *Medical image analysis*, vol. 35, pp. 303–312, 2017.

[35] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[36] H. Ravishankar, P. Sudhakar, R. Venkataramani, S. Thiruvenkadam, P. Annangi, N. Babu, and V. Vaidya, "Understanding the mechanisms of deep transfer learning for medical images," *arXiv preprint arXiv:1704.06040*, 2017.

[37] N. Dhungel, G. Carneiro, and A. P. Bradley, "Deep learning and structured prediction for the segmentation of mass in mammograms," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 605–612.

[38] J.-J. Mordang, T. Janssen, A. Bria, T. Kooi, A. Gubern-Mérida, and N. Karssemeijer, "Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks," in *International Workshop on Breast Imaging*. Springer, 2016, pp. 35–42.

[39] C. K. Ahn, C. Heo, H. Jin, and J. H. Kim, "A novel deep learning-based approach to high accuracy breast density estimation in digital mammography," in *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134. International Society for Optics and Photonics, 2017, p. 101342O.

[40] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[41] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.

[42] Y. LeCun, L. Bottou, G. B. Orr, K.-R. Müller *et al.*, "Neural networks: Tricks of the trade," *Springer Lecture Notes in Computer Sciences*, vol. 1524, no. 5-50, p. 6, 1998.

[43] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning. nature 521 (7553): 436," *Google Scholar*, 2015.

[44] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[45] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[46] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[47] Y. Taigman and M. Yang, "Marc'aurelio ranzato, and lior wolf. deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1701–1708.

[48] F. Ning, D. Delhomme, Y. LeCun, F. Piano, L. Bottou, and P. E. Barbano, "Toward automatic phenotyping of developing embryos from videos," *IEEE Transactions on Image Processing*, vol. 14, pp. 1360–1371, 2005.

[49] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3626–3633.

[50] D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[52] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[53] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[54] S. Thrun, "Is learning the n-th thing any easier than learning the first?" in *NIPS*, 1995.

[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[57] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[58] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697–8710.

[59] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[60] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2642–2651.