Research Article

Image pre-processing: enhance the performance of medical image classification using various data augmentation technique

J.Rama[1*], C.Nalini[2] and A.Kumaravel[3]

Research Scholar, Department of Computer Science, Bharath Institute of Higher Education and Research, Chennai[1]
Professor, Department of Computer Science, Bharath Institute of Higher Education and Research, Chennai[2]
Professor and Dean, School of Computing, Bharath Institute of Higher Education and Research, Chennai[3]

Abstract
The demand for techniques based on computer vision are constantly increasing due to the development of techniques for decision making pertaining to medical, social and other primary disciples of day to day life. Image processing is a subset of computer vision in which the computer vision systems make use of the image processing algorithms to carry out vision emulation for recognizing objects. This study deal with the construction of convolution neural networks (CNNs) based on deep learning. It is used for classifying chest X-ray images into two classes (Normal, Abnormal) and executed on a graphics processing unit (GPU) based high performance computing platform. Medical image classification is one of the important tasks in many medical imaging applications. Tuberculosis is a communicable disease for which early diagnosis critical for disease control. Manual screening for tuberculosis identification involves a labour-intensive task with poor sensitivity and specificity. To improve diagnosis in medical images there is in need of better classification techniques. This paper proposes CNN to classify lung X-ray images with better classification accuracy and low error rate. The data available for medical image classification problems are insufficient to train accurate and robust classifier. The data augmentation technique helps to generate more new samples from the available images using label-preserving transformations. In this paper various augmentation techniques are implemented such as horizontal flips, vertical flip, rotation (fewer angle), crops, scale right and left, are used for capturing important characteristics of medical images, and they are applied to classification function. Later little work has been done to determine which augmented strategy is best for medical image classification. Here various augmentation results are compared and evaluated to show the better augmentation techniques. It is concluded that shear lead to validation accuracies of 93% and horizontal and vertical flips gives the least accuracy of 53% of accuracy.

Keywords
Data augmentation, Flips, Rotates filters, Convolutional neural network, Shift, Scale, Shear, Tuberculosis.

1.Introduction
The medical image classification system is an important subject in the field of biotechnology. Here, the network is trained with a large amount of computation to obtain high accuracy rate [1]. Chest X-rays images (CXRs) are broadly used in identifying the abnormalities in the chest area. Automatic detecting techniques are used in most of the diagnosing process, to improve the accuracy rate of abnormality detection. The main objective of this work is to prove the range of error, loss and accuracy by using the convolution neural networks (CNNs) model for detecting tuberculosis in the chest images. Tremendous progress has been made in deep learning models for classifying medical images.

This paper proposed various augmentation strategies that can be applied in CNN to classify CXRs images. In many medical image classification problems, the data available are insufficient to train accurate and robust classifiers. To alleviate the relative scarcity of the data compared to the number of free parameters of a classifier, one popular approach is data augmentation. Data augmentation helps to transform the available samples into new samples using label-preserving transformations [2].

The main objective of this paper is to identify the better augmentation technique that can improve the performance of the medical image classifier. Therefore, multiples of new, sample data are generated by applying various augmentation techniques and the generated images are stored in a

*Author for correspondence

separate folder with the specific augmentation technique name. These images are executed with the classifier to using CNN. The proposed approach helps to identify the abnormalities in the CXRs by implementing various augmentation techniques on the dataset. As a result, it shows the appropriate classification techniques with improved performance based on increased accuracy rate with low error value. It helps to increase the possibility of diagnosis medical images using an automatic system. The automatic identification and classification of modalities from the given input image dataset are highly desired in the area of image processing. We identify the best classification method for disease identification based on CNNs techniques. Using a dataset of 2000 natural images of diseased and healthy, normal images, CNNs are trained to predict two classes either diseased or not. The deep neural network requires more data for model training to avoid under fitting [3]. The data augmentation technique generates more data from the available data set. To obtain good accuracy and robustness the option of the data augmentation strategy is important. It involves small transformations of data and those transformed data are then added to the training set for training the model [4]. The classification function is performed with the trained model to obtain significantly improved classification accuracy and robustness. Where training data is insufficient, the proposed training algorithm yields significant improvements with respect to schemes using no augmentation [5]. Data augmentation is performed on randomly selected images from an available dataset and every augmented image is grouped in a separate folder with the name of augmentation; they have further applied for training. The data being used for training plays the most important role in the classification. In spite of all available data, fetching the appropriate type of data for our experiment is an intimidating task. Moreover, the data should have good diversity; therefore, our network generalizes well during the testing phase. To overcome the problem of limited quantity and limited diversity of data, data augmentation is a technique helps to generate multiple data from the existing data. *Figure 1* shows the overview of data augmentation technique.
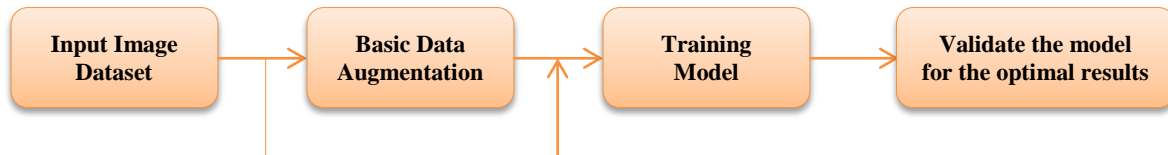

Figure 1 Overview of data augmentation technique

In this work various augmentation techniques like scaling, translation, rotation (at finer angles), horizontal flip, vertical flip, adding salt and pepper noise, Gaussian filter, shear, scale right and left and perspective transform are implemented. Various data augmentation strategies like flipping vertically and horizontally, random cropping, jitter and filters like Gaussian and salt and pepper are applied to find, the better augmentation technique for classifying medical images with improved accuracy and low loss value. Unlike in the natural image domain, where ImageNet and similar datasets provide millions of images, there are far fewer training images available in medical imaging. This dearth of training data makes it critical to explore methods such as data augmentation, which leads the data scarcity problem [2]. This study deals with the task of applying different strategies for binary image classification of normal and diseased CXRs images. It also proves that few augmentation techniques more efficient in medical image statistics more effectively than others, leading to higher training and validation accuracy. Finally, we demonstrate that smarter augmentation may result in fewer artefacts in CNN visualizations.

The rest of the paper is organized as follows. Section 2 discusses about related works. Section 3 discusses on the dataset used and preprocessing. Section 4 describes experiments and results. In section 5 concludes by summarizing the results.

2.Related works

Frid-Adar et al. (2018) [6], proposes a data augmentation method that generates synthetic medical images by using generative adversarial networks (GANs). At first the classical data augmentation techniques are implemented to enlarge the training set and then further enlarge the data size and its diversity by applying GAN techniques for synthetic data augmentation. The proposed synthetic data augmentation techniques result significantly increased to 85.7% sensitivity and 92.4% specificity.

Sebastien Wong et al. (2016) [7], proposes the purpose and advantages of augmentations. The samples obtained by augmentation of data are trained using machine learning classifier. Two approaches are implemented for creating additional training samples, the first approach is data warping, and it generates more samples by transforming the data space and over sampling. Later it is evaluated experimentally to analyse the benefits of data augmentation, using the standard MNIST handwritten digit dataset. It evaluated by three types of classification architecture, a convolutional back propagation-trained neural network, a convolutional support vector machine and a convolutional extreme learning machine classifier. The result proves that augmentation in data-space provides a greater benefit for improving performance and reducing over fitting. Fawzi et al. [5] proposed a new automatic and adaptive algorithm for transforming the samples to augmented data. The main objective is to seek small transformation that can yield a minimum classification error. Later the data augmentation scheme integrated with stochastic gradient descent (SGD) algorithm for performing the data augmentation. The experiments are conducted over two datasets and proved that the proposed method is the simplest scheme that can yield superior results to state-of-the-art methods. Salehinejad et al. [8] propose a method called generative adversarial network (GAN) using CXRs as a model medical image is implemented. A combination of real and artificial images is implied to train a deep convolutional neural network (DCNN) for detecting pathology under five classes of CXRs. It is proved experimentally that augmenting the original imbalanced dataset with a proposed method improves performance of chest pathology classification. It is explored that CNN based abnormality detection in frontal CXRs has found in the existing literature to be insufficient for making comparison of various detection techniques either due to studies reported on private datasets are not reporting the test scores in proper detail [9]. In order to overcome these difficulties, we have used the publicly available three CXRs datasets and studied the performance of various CNN architectures on different abnormalities. The classification of medical images to demonstrate the problem of data unbalance in medical imaging has been proposed [10]. The availability of the large medical image set is not feasible for manually labeling large dataset, therefore, to prepare sufficient descriptive image examples of the rare classes. So, they suggested a novel technique named unified learning of feature representation and similarity

matrix (Unified LF&SM), which are considered to be the most alike images for all categories from a huge unlabeled dataset [11]. Proposes new data augmentation techniques based on position coordinates to improve image classification performance for various imaging applications. It uses the pixel coordinates of an image for augmentation therefore to achieve better performance by resolution invariant [12]. Proposes a various augmentation strategy to find out the efficient data augmentation to capture the statistics of medical images, which can lead to new discriminative models [13]. A method to construct a large scale CXRs images with automatic diagnosing database, namely ChestX-ray8 has been proposed. The author presented Lung X-ray images comprises of 8 different thoracic diseases with their labels. They implemented data augmentation method for detecting and classifying the thoracic pathologies in the X-ray images [14]. Proposes solution to the problem in data augmentation for medical image classification. As the neural network goes deeper, the performance gets improved, but it demands more data for better performance. Medical imaging diagnosis, dealing with uncommon diseases leads to the problem of deficiency in data, which in turn leads to risking the failure for covering the vanishing gradients.

3.Dataset & pre-processing
A dataset of 909 lung X-ray images is considered for the experiment. Initially the images are divided into two folders where 490 normal images in folder namely normal and 419 tuberculosis images are saved in a folder namely abnormal. The original images from the specific folders are chosen randomly for applying augmentation, after every argumentation multiple image are generated and saved in a separate folder with the name of the corresponding augmentation technique. The total images used for the experiment increase to 2000 after augmentation in which 1000 images are abnormal images and 1000 images are normal images. Finally, the full dataset is divided into a training set and a validation set, where approximately 80% of the images are in the training set and 20% are in the validation set. The original images are accumulated from the two publicly available datasets JSRT and Shenzhen.

JSRT Dataset [15]: The data set is accumulated by the Japanese Society of Radiological Technology (JSRT). This dataset holds 247 X-rays images of the chest, out of in which 154 images have lung nodules and 93 images have no nodules. All the X-ray images are in grayscale with the size 2048 _ 2048 pixels.

Shenzhen Dataset [16]: This dataset is gathered from 'Shenzhen No.3 People's Hospital', Guangdong Medical College, and Shenzhen, China. The recorded frontal CXRs are classified into two categories: normal and tuberculosis. Out of which 336 cases having tuberculosis and 326 normal cases have been recorded from the outpatient clinics comprising a total of 662 CXRs in the dataset. *Figure 2* shows the example of normal and abnormal lung x-rays.
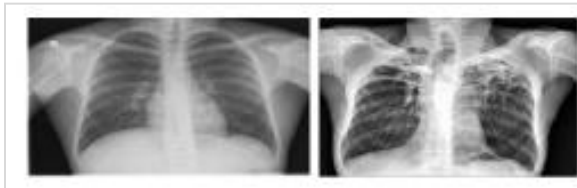

Figure 2 An example of normal lung X-rays (left) and an example of abnormal lung X-rays (right)

4.Experimental setup and results

With the help of data augmentation techniques, it is possible to generate good amount of diversity for each class of training. Image DataGenerator is frameworks that are in-built packages of python help for data augmentation. TensorFlow has TFLearn's data augmentation and MXNet has Augmenter classes. This work consists of three steps, in first step different types of augmentation are applied to the images in a dataset. Images are chosen randomly from the input dataset and nine different types of augmentations are applied. Every augmented image is saved in a separate folder. Second step, the augmented and original images are pre-processed by cropping to the size of 200×200 and later every augmented data sets are spitted into training and validation sets. Second, for each image the training set, we perform one of eight augmentations. Third step the pre-processed images are trained independently on the eight uniquely augmented sets. Finally, we evaluate performance of the trained CNN by measuring training and validation accuracy of normal and abnormal classification as well as qualitatively assessing visualizations generated from the CNN. As we have used an equal number of images in the dataset of each class (Normal, Abnormal) accuracy and average error loss are used as an evaluation measure. Accuracy is measured can be a better measure for computing large balanced class.

4.1Construction of convolutional neural network

Step 1: Defining an input function - The lung images of size 200×200 pixels are converted as a TFrecords and fed as input setting batch size as 100 and the

10

buffer size as 1000. This returns tensor of an image and its corresponding label.

Step 2: Defining model - Three convolutional layers are defined and implemented with 16 channels and 3×3 filters having concatenated rectified linear units (CReLU) as activation function. Following the convolutional layer three pooling layers (PL1, PL2, PL3) are defined with max pool function having stride value as 1 for all the dimensions and activation function as CReLU. Finally, two fully connected layers (FCL1, FCL2). The last fully connected layer FCL2 consists of two nodes equal to the size of classes.

Step 3: Evaluation metrics are calculated by computing classification accuracy.

Step 4: The created model trained between 100 to1000 iterations. The maximum accuracy obtained is considered as a result.

Step 5: The performances are evaluated on the validation set. Therefore, the convolutional network predicts any of the two classes of the images.

4.2Steps involved in data augmentation

Step 1: Load and format the input image data.

Step 2: Image generator is created using ImageDataGenerator function.

Step 3: Changes apply to the parameters of the image generator function to obtain a new augmented image.

Step 4: keras.Preprocessing library is used for exporting the generated augmented images to their corresponding folders.

Step 5: Repeat step3 and step4 for various type augmentations.

Step 6: Train and validate the proposed CNN model augmentations before training the models. We utilize nine augmentation strategies to generate nine new training sets. Each new training set is simply the original training images in addition to the training images augmented by one of the techniques below. Each augmentation is as follows:

We utilize nine augmentation strategies to generate nine new training sets. Each new training set is simply the original training images in addition to the training images augmented by one of the techniques below. Each augmentation is as follows:

Flips (Horizontal & Vertical): Horizontal and vertical flips are applied to the randomly chosen images from the original training set. Those flipped images are saved separately in two different folders namely horizontal flips and vertical flips respectively. Though both the flips can capture unique property of medical images, it is considered.

Rotation: Rotations are applied to the randomly chosen image from the original training set. Rotations are applied with fewer angles, where angle ranges between 20 degrees to 90 degrees are applied to the randomly selected images. Those rotated images are saved separately in folders namely rotation.

Gaussian Noise: Images are randomly selected from the original dataset. Gaussian noise is applied to the selected images, therefore to reduce the unwanted noise by blurring the edges and reduces the contrast and they are saved in a separate folder namely Gaussian.

Jittering: Images are randomly selected from the original dataset. Then, for each chosen image applied with the least amount of contrast ranges between + or − 1 to 4 and they are saved in a separate folder namely Jitter.

Shift (Height and Width): Images are randomly selected from the original dataset. Then, for each chosen image are shifted with the value ranges between 0.3 and 0.5 respectively, and they are saved in a separate folder namely shift.

Shears: Images are randomly selected from the original dataset, then for each chosen image are horizontal and vertical shearing with the value ranges between 0.2 and 0.5 respectively and they are saved in a separate folder namely shear. Finally, each chosen image is sheared.

Scaling: Images are randomly selected from the original dataset. Then for each chosen image horizontal and vertical shearing with the value ranges between 0.10 and 0.25 respectively, and they are saved in separate folder namely scale.

The diagnosing element in the image may be tiny or large. Scaling inward reduces the image size helps in diagnosing abnormalities that lies closer the boundary. Finally, each chosen image is scaled.

Zoom: Images are randomly selected from the original dataset, then for each chosen image are zoom with the value ranges between -1 and 1 respectively and they are saved in a separate folder namely zoom. Finally, each chosen image is zoomed.

Adding Salt and Pepper noise: When NN learns high frequency features it cause over fitting. There are few irrelevant features that are removed by introducing Gaussian noise, which has zero mean data point, basically at all frequencies that leads to distorting the high frequency features. Adjusting the input images by just adding the right amount of noise can enhance the learning capability. The mean and standard deviation values are set between 0 to 0, 1 for adding noise to the input images and stored in separate folder namely noise.

All the original and augmented images in the training and validation set are cropped to 200×200 pixels. Then, the images in training sets are trained using CNNs.
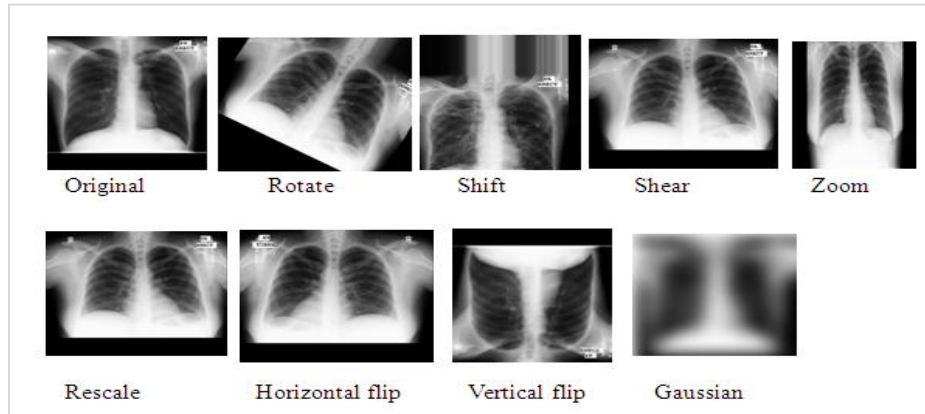
Figure 3 shows the sample augmented image.

Figure 3 Sample augmented image

All the methods are implemented, when the data augmented and the activation function is set as CReLU and the number of convolution layers and pooling layers are three. As depicted in *Figure 4* the maximum accuracy rate obtained is 94.4% and the subsequent average loss is 8.5% when augmentation technique is shear. All the methods are implemented, when the data augmented and the activation function is set as CReLU and the number of convolution layers and pooling layers are three. As depicted in *Figure 4* the maximum accuracy rate obtained is 96.8% and the subsequent average loss is 8.3% when augmentation technique is shear. The result determines if the extent to which an augmentation captures information about the original training set affects the training and validation accuracies.

11

Because the original training set is representative of the validation set, this effect gives us some intuition on how the model might generalize given some augmentation on the training set.
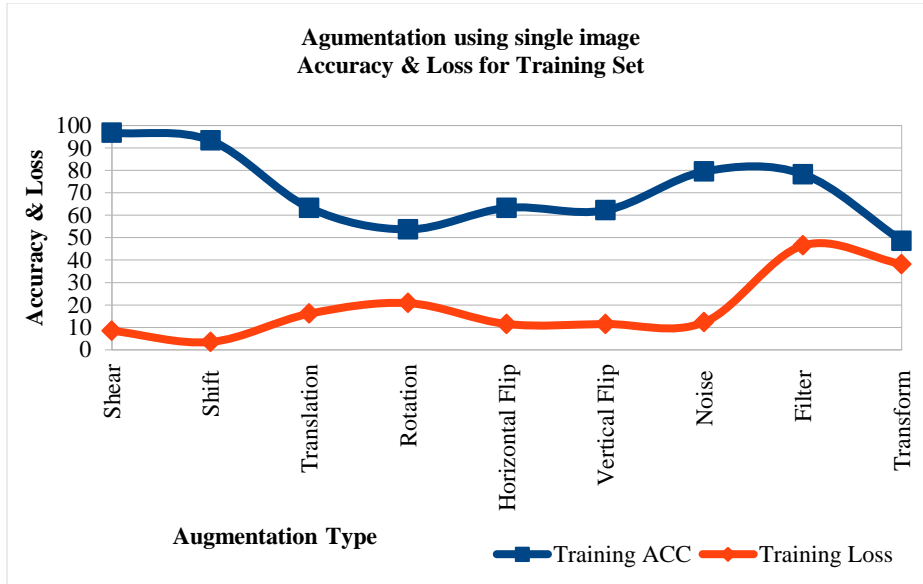
Agumentation using single image Accuracy & Loss for Training Set

Figure 4 Accuracy and loss value obtained by the training set

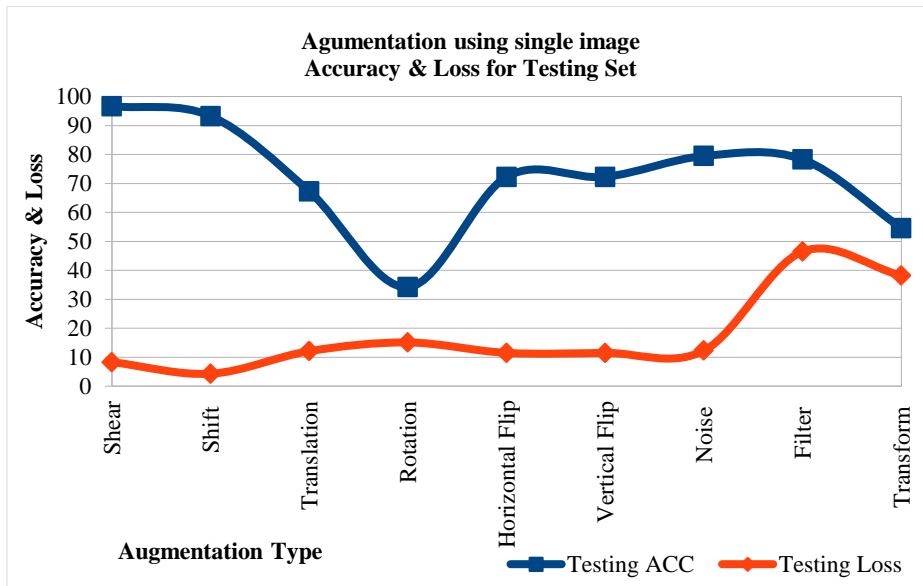**Agumentation using single image Accuracy & Loss for Testing Set**

Figure 5 Accuracy and loss value obtained by testing dataset for various augmented images

5.Discussion

The accuracy and loss value of various augmented data are demonstrated as the graph in *Figure 4* and *Figure 5*, the accuracies for the shear and shift (height and width) augmentation sets are higher value as 96.8% for testing test, when comparing all other augmentation set. Specifically, the augmentations that resulted in zoom augmentation obtain the lowest accuracy value of around 48.6%. Also, in general, the noise and Gaussian filter augmentations, resulting in accuracy 79.5% and 78.3%, respectively, have more variance in their validation accuracies over time compared to augmentations resulting in higher information. Considering the shear and shift augmentations, which result in mean images with relatively low mutual information, but high classification accuracies, there is a perceptive justification for the case, namely, shift and shears do

12

maintain many of the image information in the original training set, leading to high classification accuracy, but artefacts in the augmented images may affect the main image and lead to a lower mutual information. For example, if a mass appears on the edge of a chest, then the 200 by 200 crop of the original image will preserve the 0 pixels that appear when shift right and left image. Usually, if the abnormality is near the centre, as is the case for the majority of the training images, then the 0 pixels will be cropped away. These slight artefacts may decrease the overall information, even though the high classification accuracies suggest that shift and shears preserve mammography image statistics. To solidify the correlation between mutual information and classification accuracy. Furthermore, we see that the augmentation strategy used has a significant effect on the type of mass that is generated, as shown by the images in *Figure 4* and *Figure 5*. In general, we see that the augmentation determines the type of artefact that predominates the mass generation. For example, the artefacts that appear in the generated image from the CNN trained on the rotations set have circular patterns. We see a similar phenomenon in the 'flips' generated image, where we see several abnormalities that look, they have been flipped. These results obtained by the execution are comparable with existing [2] work obtained of maximum accuracy value of 88% for rotating augmentation using testing set and proved that the proposed method yields the better accuracy of 96%.

6.Conclusion
To improve diagnosis in medical images there is in need of better classification techniques. This paper addresses the issue of medical image insufficiency for classification problem. The data augmentation technique helps to generate more new samples from the available images using label-preserving transformations. In this work various augmentation techniques are implemented such as horizontal flips, vertical flip, rotation (fewer angle), crops, scale right and left, are used for capturing important characteristics of medical images, and applied to classification function. Later little work has been done to determine which augmented strategies are best for medical image classification. Here various augmentation results are compared and evaluated to show the better augmentation techniques. It is concluded that shear lead to validation accuracies of 93% and horizontal and vertical flips gives the least accuracy of 53% of accuracy.

References
[1] Ciompi F, Chung K, Van Riel SJ, Setio AA, Gerke PK, Jacobs C, et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. Scientific reports. 2017; 7:46479.
[2] Hussain Z, Gimenez F, Yi D, Rubin D. Differential data augmentation techniques for medical imaging classification tasks. In AMIA annual symposium proceedings 2017 (p. 979). American Medical Informatics Association.
[3] Razzak MI, Naz S, Zaib A. Deep learning for medical image processing: overview, challenges and the future. In Classification in BioApps 2018 (pp. 323-50). Springer, Cham.
[4] Inoue H. Data augmentation by pairing samples for images classification. arXiv preprint arXiv:1801.02929. 2018.
[5] Fawzi A, Samulowitz H, Turaga D, Frossard P. Adaptive data augmentation for image classification. In international conference on image processing 2016 (pp. 3688-92). IEEE.
[6] Frid-Adar M, Klang E, Amitai M, Goldberger J, Greenspan H. Synthetic data augmentation using GAN for improved liver lesion classification. In international symposium on biomedical imaging 2018 (pp. 289-93). IEEE.
[7] Wong SC, Gatt A, Stamatescu V, McDonnell MD. Understanding data augmentation for classification: when to warp? In international conference on digital image computing: techniques and applications 2016 (pp. 1-6). IEEE.
[8] Salehinejad H, Valaee S, Dowdell T, Colak E, Barfett J. Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks. In international conference on acoustics, speech and signal processing 2018 (pp. 990-4). IEEE.
[9] Kumar A, Wang YY, Liu KC, Tsai IC, Huang CC, Hung N. Distinguishing normal and pulmonary edema chest x-ray using Gabor filter and SVM. In international symposium on bioelectronics and bioinformatics 2014 (pp. 1-4). IEEE.
[10] Zhang J, Xia Y, Wu Q, Xie Y. Classification of medical images and illustrations in the biomedical literature using synergic deep learning. arXiv preprint arXiv:1706.09092. 2017.
[11] Kori A, Krishnamurthi G, Srinivasan B. Enhanced Image Classification with data augmentation using position coordinates. arXiv preprint arXiv:1802.02183. 2018.
[12] Gupta A, Chopra S, Ledig C. Generative bone lesions synthesis for data augmentation in X-ray.2018.

J.Rama et al.

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

[14] https://cdn.ymaws.com/siim.org/resource/resmgr/siim 2018/abstracts/18posters-Borstelmann.pdf. Accessed 12 November 2018.

[15] Shiraishi J, Katsuragawa S, Ikezoe J, Matsumoto T, Kobayashi T, Komatsu KI, et al. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. American Journal of Roentgenology. 2000; 174(1):71-4.

[16] Jaeger S, Candemir S, Antani S, Wáng YX, Lu PX, Thoma G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quantitative Imaging in Medicine and Surgery. 2014; 4(6):475-77.

Miss A. Rama is working as an Assistant Professor in the Department of Information Technology, Bharath University, Chennai. Her research area includes Visual Computing and Machine Learning.
Email: rama_j1@yahoo.com

Dr. Nalini Chidambaram is working as a Professor in the Department of Computer Science and Engineering, Bharath University, Chennai. Her research area includes Data Mining, Social Networking and Image Processing.

Dr Kumaravel is working as a Professor and Head in the Department of Information Technology, Bharath University, Chennai. His research interest includes Data Mining, Programming Language and Software Engineering, Visual Computing, Machine Learning, Pervasive Computing and Knowledge Engineering. He is a Life Member of the Computer Society of India (CSI).