International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 264-280 www.ijaem.net ISSN: 2395-5252

Application of Gan in Data Set Generation for Detection of Melanoma Skin Cancer

SalihuAlhassan Libata[1], Dr. Haasan Suru[2], Mal. Abubakar Ahmed Aliero[3]

Department of Information Technology, Continental Transfert Technique Limited, Abuja, Nigeria[1]
Department of Computer Science, Kebbi State University of Science and Technology Aliero, Nigeria[2]
Department of Internal Medicine, Federal Medical Center, BirninKebbi,Nigeria[3]

Submitted: 01-03-2022 Revised: 10-03-2022 Accepted: 12-03-2022

ABSTRACT: Melanoma is the most dangerous type of skin cancer. Early diagnosis is crucial to increase the survival rate of patients due to the possibility of metastasis. Automated skin lesion detection can play an essential role by reaching people that do not have access to a specialist. Over the past few years, Deep learning has been widely used in medical imaging for classification and segmentation and has been successful in providing better diagnostic accuracy. State of the art deep learning algorithms are built using neural networks arranged in layers where first layer extracts basic information of images like edges, colors etc so that the output of one layer is fed as input to the next consecutive layers. Thus, increasing the complexity of learning with increase of layers. In comparison with the traditional machine learning algorithms, deep learning has many advantages and is an automatic process. However, it requires large scale annotated data for better performance and is thus constrained by limited size of available public datasets. To overcome data constraints, the objective of this dissertation is to solve the problems that arise by having limited datasets. In this work, images are augmented to increase the dataset size using GAN based augmentation. GAN based augmentation is based on neural networks, where two neural networks compete against each other to produce visually realistic synthetic images. For the classification, CNN and an ensemble of CNNs are trained on the final enlarged training dataset comprising original images and synthetic images to boost performance.

Our method generates high-resolution clinically-meaningful skin lesion images that when compound our classification model's training dataset, consistently improved the performance in different scenarios, for distinct datasets. We also investigate how our classification models perceived the synthetic samples and how they can aid the model's generalization.

Keywords: Melanoma Skin Cancer, GAN Technique, skin image segmentation, skin lesion classification, machine learning, deep learning, Convolutional Neural Network (CNN).

I. INTRODUCTION

According to the American Cancer Society, [1] Cancer is one of the leading causes of death in modern times with a death rate of 1 death per 6 people. there were seventeen (17) million new cancer cases and nine point five (9.5) million deaths due to cancer in the year 2018. Skin cancer is one of the deadliest cancers in the United States. Out of the two major types of skin cancer, melanoma is fatal and has estimated five-year survival rate of about 99% if detected early and 20% if detected late. [2] In the year 2018, melanoma was ranked ninth amongst ten major cancers in the United States and in 2021 was estimated to have more than 207,390 new cases of melanoma, about 106,110 of which will be invasive (that is penetrating to the skin second layer) with estimated deaths of 7,180 (4600 men and 2580 women) [3]. Only 20-30% of melanoma is found in existing moles, while 70-80% arises on normal looking skin. Melanoma occurs when a melanocyte (melanin pigment forming cell) starts multiplying in an uncontrollable way and the multiplication results in formation of malignant tumors. Such malignant tumors are characterized by features like asymmetry, irregular borders, multiple colors, diameter greater than 6mm and c lesion. asymmetry, borders, colours, Diameter and Enlarging are usually represented by an acronym (ABCDE). [4]Out of all the listed features, colour and structure play vital role in the diagnosis of melanoma. The structure is characterized by pigment distribution, symmetry or asymmetry, homogeny or heterogeny, skin surface keratin, regular or irregular vascular morphology, presence or absence of ulceration and border of lesion, and in case of colour, black, brown, yellow, white and grey.

These features are examined to diagnose melanoma in early stage. If diagnosed early, melanoma can be treated. Many algorithms in medical imaging have been applied to improve performance for an accurate and early diagnosis.

Medical imaging is a technique of representing the interior or exterior of the body in visual form, commonly referred to as "imaging modality" by using radiations, radio frequencies, sound waves Et cetera for diagnosis. Interiors of the body are widely studied by using imaging modalities like Computed Tomography (CT) scan, X-ray, Positron Emission Tomography (PET) and Magnetic Resonance Imaging (MRI). Skin, being the exterior of a body is captured by using dermatoscopy. Dermatoscopy is one of the imaging modalities that uses polarized light to capture skin images and is done using an instrument called digital epiluminescencedermatoscope. Research shows that the Dermatoscopy increased diagnostic accuracy by 10-27% when compared to the human experts, who had been assessing melanoma on basis of their knowledge and naked eye examination [5]. Over the past few decades, different areas in computer science are emerging to deliver promising results to aid the diagnostic process.

Artificial Intelligence (AI), being one such fields, is directed towards automating the diagnostic process by developing complex algorithms and has the potential to outperform human experts.

Research conducted by a team from Germany, the United States and France "Man against machine" European Society for Medical Oncology, demonstrated effectiveness of AI system by using an AI system to detect and classify malignant and benign skin lesions by feeding more than 100,000 images of skin lesion into Convolutional Neural Network (CNN). The AI system was able to classify with 95% accuracy while human experts were able to perform with 86.6% accuracy. Computer vision is one subset where AI automates the process by taking input data in the form of images. The AI system utilizes an ensemble of techniques where algorithms learn iteratively from data. This learning process is defined as Machine Learning (ML).

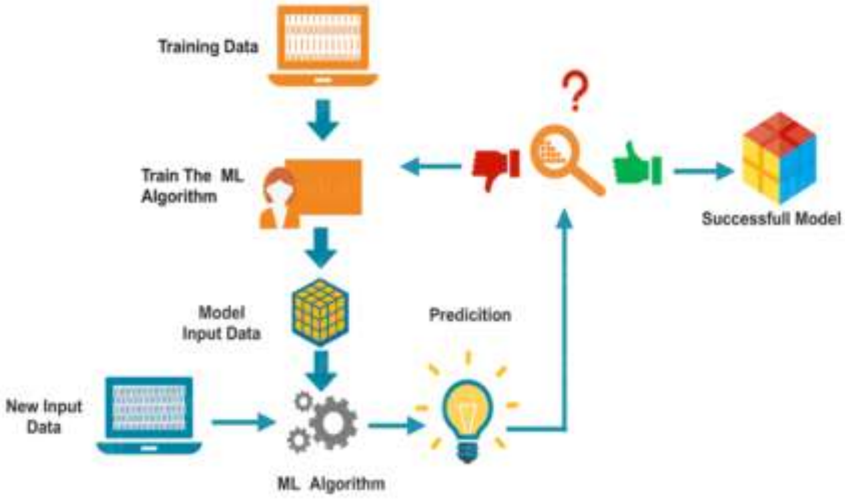

Fig. 1 Machine learning

When we feed an input data with its label, ML algorithm tries to learn features associated with labels and such learning is typically known as supervised learning.

A. SUPERVISED LEARNING

Supervised learning applies a tagged data set containing input values and expected output values. When training AI with supervised learning, we need to enter a value for it and tell it the expected output value. If the output value generated by the AI is incorrect, it will adjust its calculation. This process will iterate as the dataset is updated until the AI no longer makes mistakes.

A typical application for supervised learning is the weather forecast AI application. AI uses historical data to learn how to predict weather. The training data includes input values (air pressure, humidity, wind speed, etc.) and output values (temperature, etc.).

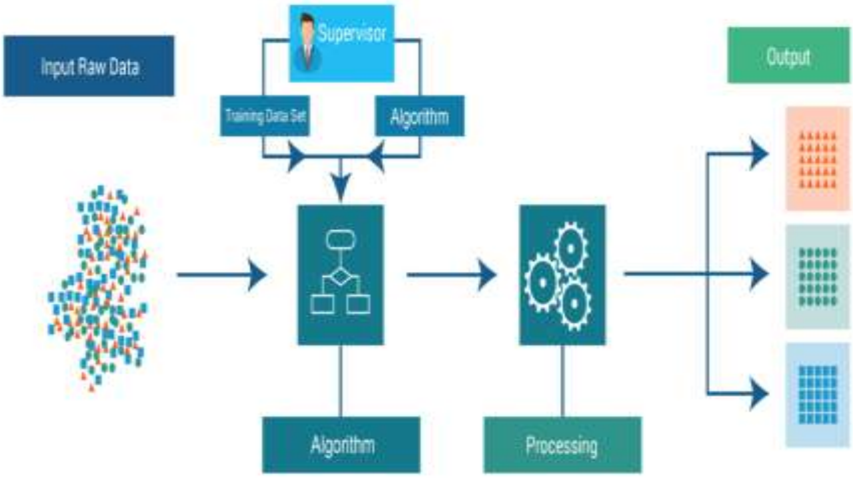

Fig. 2 Supervised Machine learning

B. UNSUPERVISED MACHINE LEARNING

Unsupervised learning is the use of data sets without specific structures. When training AI using unsupervised learning methods, you need to have AI classify the data. An example of an application for unsupervised learning is to predict consumer behavior for e-commerce sites. AI does not take advantage of the tagged input and output value data sets. Instead, it classifies the input data itself, allowing the site to know what customers like to buy.

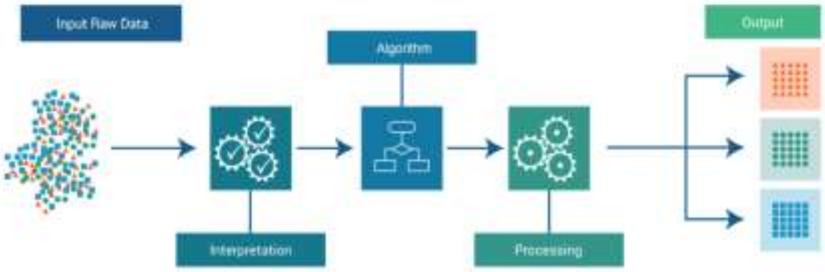

Fig. 3 Unsupervised learning

Since an algorithm tries to learn from data with no labels, data in medical imaging are often annotated or labeled by human experts. Therefore, most of medical diagnostic problems use supervised learning. After training the labeled data, efficiency of algorithm is verified by testing it with unseen test data. If an algorithm shows similar performance and can correctly identify test data, then algorithm is validated and can be regarded as "efficient".

Combination of classical hand engineered feature extraction with traditional machine learning algorithms like Support Vector Machine, Random forest etc. have shown efficient results in melanoma diagnostics [6], [7]. At the same time, traditional ML algorithms have limitations to process the data in its raw form and thus a prerequisite is a carefully designed of feature extractor that is able to represent features of data compatible for classifier to learn. To overcome the challenges faced by traditional ML algorithms, deep learning has emerged as a subset of ML. Deep learning method introduced representation learning, allowing automatic feature extraction, detection or classification simply by feeding in raw data. This required little hand engineering and dramatically improved the classification process in computer vision [8]. Convolutional Neural Network (CNN) is considered as the most popular deep learning technology in computer vision [8]. It has layers of neural network arranged in hierarchical manner with the output of one layer fed as input to the next layer. Here, the first layer extracts basic information of images like edges,colour Et cetera and complexity of extraction keep on increasing with the number of layers. With enough and proper training, CNNs are able to learn and extract features automatically. In computer vision application, CNN takes in an input image, assigns importance by updating weights and biases to different features in the image and is able to classify one from another. Understanding the role and efficacy of an efficient deep learning method over traditional ML algorithms for medical diagnostics, it is highly desirable to build an efficient deep learning model for melanoma detection as early detection can increase the chances of survival by 99%.

II. LITERATURE REVIEW

Melanoma is the most dangerous form of skin cancer. It causes the most deaths, representing about 1% of all skin cancers in the United States. According to the American Cancer Society, the crucial point for treating melanoma is early detection. The estimated 5-year survival rate of diagnosed patients rises from 15%, if detected in its latest stage, to over 97%, if detected in its earliest stages.

Cancer develops when cells in the body begin to proliferate uncontrollably. Metastasizing means that cancerous cells may form in practically any place of the body and spread. In this regard, the uncontrolled proliferation of abnormal skin cells is referred to as skin cancer. Uncorrected DNA damage to skin cells, most typically produced by UV radiation from the sun or tanning beds,]creates mutations, or genetic flaws, that cause skin cells to reproduce rapidly and produce malignant tumors.

According to [9], professional dermatologists have achieved 90% sensitivity and 59% specificity in the identification of skin lesions. Around the same time, the statistics for less qualified doctors indicated a significant decline for general practitioners to about62–63%.

Deep learning is an intuitive process whose complexity of learning increases with the increase in the number of layers. Due to its high performance, it is regarded as a mature application for medical diagnostics [10]. In recent times, deep learning has contributed significantly for skin lesion classification problems [11], [12]. However, limited data set creates tougher environment for the potential groundbreaking research in medical diagnostics with deep learning. One reason is dependency of the deep learning algorithm on training data size as it requires millions of parameters and large amount of labeled data to learn [13]. When limited data is used to train deep learning model, it uses large amount of its resources to train the model, creating over fitting issues. Over fitting issue refer to a model's incapability to generalize on unseen data. A large numbers of research have been done to overcome challenges imposed by limited data on the training of deep learning models. It includes techniques like augmentation [14], transfer learning [13]and ensemble of classifiers [15]. A visual inspection by a dermatologist of the suspicious skin region is the first stepin diagnosing a malignant lesion. A correct diagnosis is essential because certain types oflesions have similarities; moreover, the accuracy of the Computer-Aided System (CAD) isclose to the experienced dermatologist's diagnosis. [16]

III. OVERVIEW OF GAN TECHNIQUE

GAN is the deep convolution neural network that was defined by a team of research workers under the guidance of Ian Good fellow. GAN has two competing neural network model. One uses the noise as input and generates samples (and so named generator). The second model which is known as discriminator receives samples from the generator and the training data. According to game theory, the generator is trained to produce an image that looks like a real image, whereas the discriminator is learning to discriminate perfectly from generated data to actual data.

The GAN is trained similarly to the Minimax algorithm from Game Theory and the two networks try to achieve Nash Equilibrium with respect to each other. In medical image synthesis, MRI imaging is obtained from CT imaging. The generator produces the CT image and the discriminator is trained to differentiate the generated CT image with reference to the desired input (CT image) given to it. GAN is the perfect model for generating images and images with better resolution compared to the images generated by MLE (maximum likelihood Estimation) algorithm.

A. STRUCTURAL VARIANTS

GAN is the deep neural framework that contains two models namely the generative (G) and discriminative (D) model. G produces fake samples looks like training sample from the latent variable z, whereas faux samples from G is given to the D. Discriminator find the difference between real data and data generated from G.G tries to satisfy the discriminator with the faux sample, whenever the faux data is found by the discriminator the error is back propagated to the generator. This adversarial learning is formulated in equation (1)

$\min GmaxD\,V(G,D)=\min GmaxD\{Ea\sim pdata[\log D(a)]+Ez\sim pzz\,[\log(1-D(G(z))]\}$………..(1)

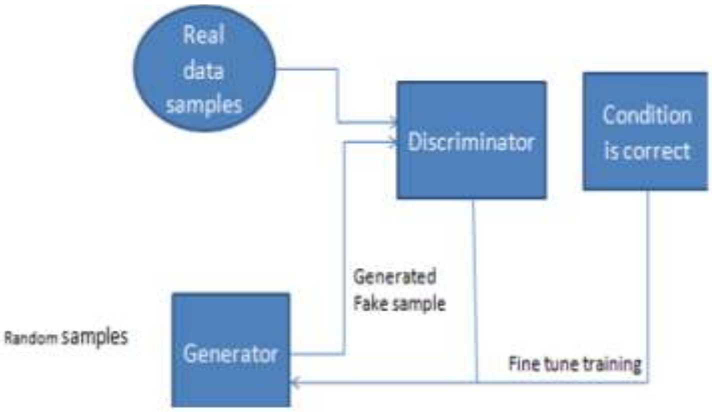

Fig. 4GAN Architecture

Whereas pdata (a)-distribution of real data in the data spaceA pz(z)-distributor generator on the latent space Z.

The generator and discriminator neural model contains the differential function where the weights and bias of two models can be modified to adjust the probability density function through the back propagation algorithm. Generator satisfies the discriminator with faux data when the PDF of input(pdata(a)) is equal to the PDF of the data created by the generator(pg(a)) [17]. In this instance it is difficult to differentiate faux and real data hence the discriminator produces the output as 0.5. It shows that D gets confused. The objective of the discriminator is to maximize the two terms log (1-D (G (z))) and log (G(a)) i.e. D(z)=0 and D(a)=1 to correctly classify the faux and real data.

The best or optimal generator is given as:

DG*(a) =$(a)pg(a)+pdata$(a)……….(2)

Here b=0 is the generated data and b=1 is the real data (b is assumed to be the output) and it is assumed that the probability of both generated and real data are equal. This condition represents that the relative behavior of the two distributions, [18], is used by the GAN to pass the collection of data to the generator for generating actual samples. GAN also measures the discriminator discrepancy.

IV. CONVOLUTIONAL NEURAL NETWORK (CNN)-BASED SKIN CANCER DETECTION TECHNIQUES

A convolution neural network is an essential type of deep neural network, which is effectively being used in computer vision. It is used for classifying images, assembling a group of input images, and performing image recognition. CNN is a fantastic tool for collecting and learning global data as well as local data by gathering more straightforward features such as curves and edges to produce complex features such as shapes and corners. [19]. CNN's hidden layers consist of convolution layers, nonlinear pooling layers, and fully connected layers [20]. CNN can contain multiple convolution layers that are followed by several fully connected layers. Three major types of layers involved in making CNN are convolution layers, pooling layers, and full-connected layers [21]. The basic architecture of a CNN is presented in (Figure 5)

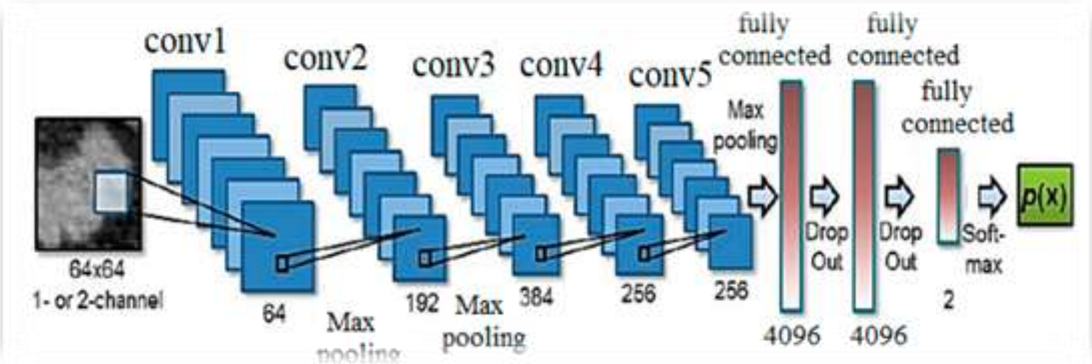

Fig. 5 Basic CNN Architecture.

CNN-based automated deep learning algorithms have achieved remarkable performance in the detection, segmentation, and classification operations of medical imaging [22]. [23], proposed a very deep CNN for melanoma detection. A fully convolutional residual network (FCRN) having 16 residual blocks was used in the segmentation process to improve performance. The proposed technique used an average of both SVM and softmax classifier for classification. It showed 85.5% accuracy in melanoma classification with segmentation and 82.8% without segmentation. [11], proposed a multi-scale CNN using an inception v3 deep neural network that was trained on an ImageNet dataset. For skin cancer classification, the pre-trained inception v3 was further fined-tuned on two resolution scales of input lesion images: coarse-scale and finer scale. The coarse-scale was used to capture shape characteristics as well as overall contextual information of lesions. In contrast, the finer scale gathered textual detail of lesion for differentiation between various types of skin lesions. A research conducted by [24], proposed a technique to extract deep features from various well established and pre-trained deep CNNs for skin lesions classification. PretrainedAlexNet, ResNet-18 and VGG16 were used as deep-feature generators, and then a multi-class SVM classifier was trained on these generated features. Finally, the classifier results were fused to perform classification. The proposed system was evaluated on the ISIC 2017 dataset and showed 97.55% and 83.83% area under the curve (AUC) performance for seborrheic keratosis (SK) and melanoma classification. A deep CNN architecture based on pre-trained ResNet-152 was proposed to classify 12 different kinds of skin lesions, [24]. Initially, it was trained on 3797 lesion images; however, later, 29-times augmentation was applied based on lighting positions and scale transformations. The proposed technique provided an AUC value of 0.99 for the classification of hemangioma lesion, pyogenic granuloma (PG) lesion, and intraepithelial carcinoma (IC) skin lesions. A technique for the classification of four different types of skin lesion images was proposed by [25]. A pre-trained deep CNN named AlexNet was used for feature extraction, after which error-correcting output coding SVM worked as a classifier. The proposed system produced the highest scores of the average sensitivity, specificity, and accuracy for SCC, actinic keratosis (AK), and BCC: 95.1%, 98.9%, and 94.17%, respectively. [26], proposed a pre-trained deep CNN architecture VGG-16 with a final three fine-tuned layers and five convolutional blocks. The proposed VCG-16 model is represented in Figure 6. VCG-16 models showed 78% accuracy for the classification of lesion images as melanoma skin cancer. A deep CNN-based system was proposed to detect the borders of skin lesions in images. The deep learning model was trained on 1200 normal skin images and 400 images of skin lesions. The proposed system classified the input images into two main classes, normal skin image and lesion image, with 86.67% accuracy.

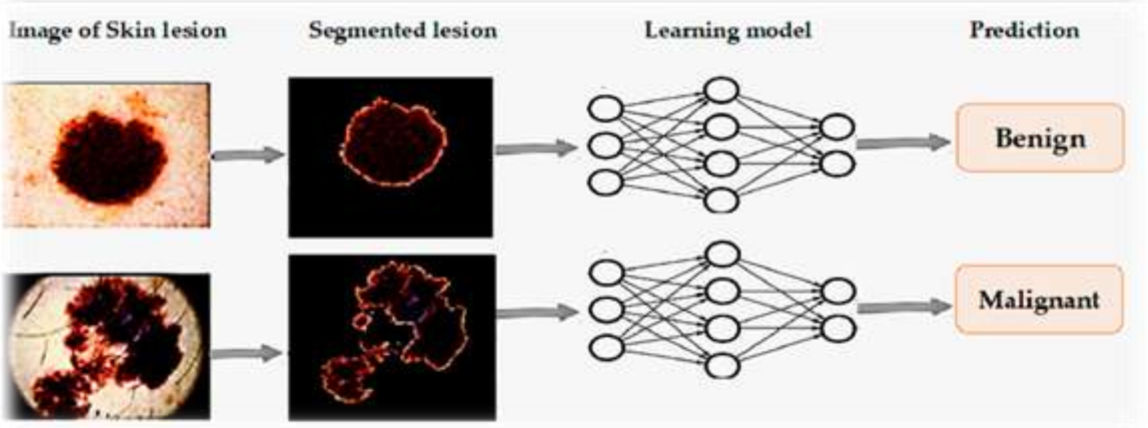

Fig. 6 Skin cancer diagnosis using CNN

V. THE PROPOSED METHOD

In super resolution GAN (SRGAN), input to the generator is a conditional input 'c' concatenated with random noise. The objective function in the original GAN was defined as a min-max with optimization goal to reach the Nash Equilibrium [27], [28], which led to vanishing gradient problem [29]. SRGAN proposed to use wasserstein as an objective function. In comparison to the original GAN, SRGAN trains its discriminator/critic using wasserstein distance instead of Jensen-Shannon divergence to calculate the distance between the estimate and actual data distribution. Using wasserstein distance as a cost function, the training process demonstrated better performance with no sign of mode collapse and in comparison with the old works, wasserstein loss function reflected image quality in a better way, [30]. The pre-trained generator network was trained using the content loss which is a simple L2-loss based on the difference of a generated and a target image CNN feature maps [30]. Here, the discriminator is referred to as a critic as it doesn't give output in the form of 0 or 1 but instead gives numerical value and is trained using wasserstein GAN (WGAN) loss with addition of gradient penalty to avoid hyper tuning of parameters and to restore the finer texture details [31]. Thus, SRGAN is represented by two loss functions that are defined at two different stages.

In proposed approach, the process of image augmentation takes place in two stages. In the first step, the generator network pre-trained on Imagenet begins to generate the original synthetic samples. It occurs when the loss functions of generator and discriminator converges. In the second step, as per the core concept of GAN, the discriminator and the generator are in adversary with each other. The discriminator tries to distinguish fake synthetic images from the generator and adversarial training continues till discriminator takes samples generated by generator as real samples. Finally, these original synthetic samples that are able to fool discriminator are taken as final synthetic images and are used as training images for CNN classifier. In our proposed method, we added blur to original images such as gaussian blur, mean blur, median blur, bilateral blur, motion blur, mixture-of-motion-and-median blur and bright blur and generated a blur dataset. We generated 1000 images for each class using each blur. Blurred images are input to the generator in SRGAN while blur kernel is opaque to the generator.

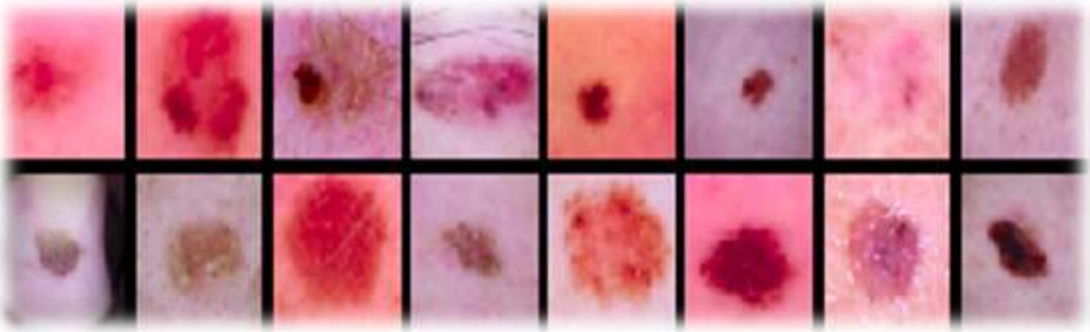

Fig. 7 Example of Real images

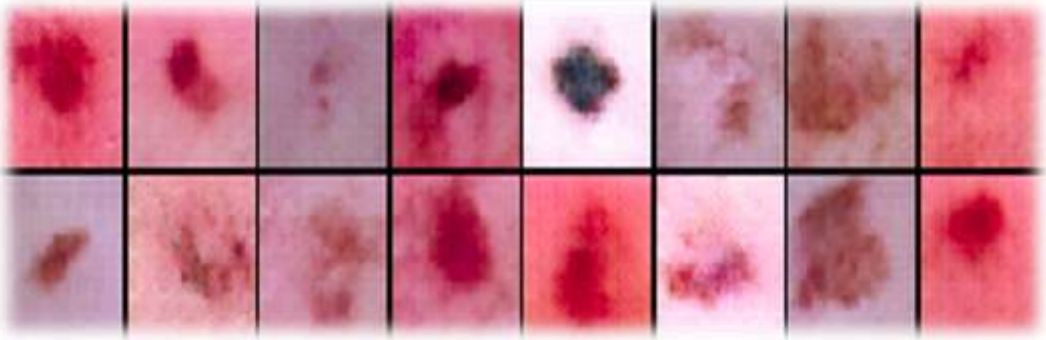

Fig. 8 Example of Generated images

IV. COMPONENTS OF THE PROPOSED METHOD

SRGAN is able to solve vanishing gradient problem by implementing Wasserstein loss with gradient penalty, moreover it is able to overcome mode collapse problem by implementing content loss [30]. Training process is similar to the original GAN where loss is calculated at discriminator on discriminator's output and is backpropagated to generator. However, content loss has been introduced at generator's output to ensure the deblurring process. Similar to the normal GAN, there are four main components in

SRGAN are discussed in detail in following subsections:

A. THE DISCRIMINATOR

It is a convolutional neural network which outputs a single value. The objective of this network is to determine whether the input image is fake or real. To build the full model, the generator's output is connected to the discriminator's input. While training the discriminator, we pause the training of the generator and start the training of the discriminator using wasserstein loss on images generated by the generator. Images generated by the generator are used as the negative examples during training process. During training, the discriminator classifies the data generated by the generator into real or fake based on its loss function which is a wasserstein loss. For each misclassified case, the discriminator loss penalizes the discriminator and then, the discriminator updates its weights via back propagation into the discriminator network. It is same as shown in figure 3.2. Gradient penalty has been added to Wasserstein loss which is robust to the generator architecture. This is beneficial as it removes the step of hyper parameter tuning to some extent.

The architecture of discriminator is convolutional. In our work, we defined discriminator/ critic network during training phase which is wasserstein GAN with gradient penalty. Architecture of critic network takes input image of size 64x64x3. Network consists of four convolutional layers with a kernel size of 4x4 and a fully connected layer. Batch normalization is used after second and third convolution. All convolutional layers except last layers are followed by Leaky ReLU activation of 0.2. In the final layer sigmoid activation is used. In first two steps, we build generator and discriminator architecture and then, final model is constructed by having direct feedback on generator's output.

B. THE GENERATOR

It is a neural network based on unsupervised learning. It takes in an input of blur images and tries to generate the real sharp images. Here, this network is based on nine resnet blocks. The resnet blocks are also called residual blocks. Unlike in the traditional neural networks where the output of each layer is fed into the next layer, in residual blocks the output of one layer is fed into the next layer and also, into layers about 2-3 hops away. This connection is commonly referred to as skip connections.

The residual blocks learned the true distribution of original images by learning residue (R(x)) which is defined as the difference between generator's input (x) and true distribution of original images (H(x)) and is written as R(x) = H(x) - x . In the original GAN, the generator layers were trying to learn H(x) but here, layers are learning the residue. The use of residue makes it easier to backpropagate the gradients into the generator because of the available options of skip connections. Training of generator is done by using backpropagation. Generator generates synthetic images and feeds in the generated images to the discriminator. When the discriminator classifies generated images as fake, the generator is then trained on content loss and at this time, the discriminator's training is paused. The generator then updates its weight to produce the sample better than the one in previous iteration. Training process is like the one in (figure 3.3). the research work followed standard architecture given by [30], for the generator. The architecture of generator consists of (i) two strided convolution blocks (strides = 1/2), (ii) nine residual blocks each residual block further consists of convolution layer, batch normalization(BN) and ReLU activation with dropout of 0.5 and (iii) two transposed convolution blocks. After the final activation layer, connection is added from the input to the output and further, normalized output by dividing it by 2.

C. THE LOSS FUNCTION

In the section 3.1.2, we observed that original GANs do not converge due to min max objective function which uses minimization of JSD distance. Using Wasserstein distance instead of JSD to measure difference between generated and original data distribution guarantees convergence and differentiability. The wasserstein distance takes mean of differences between two images (restored and original) [32], and calculates minimum cost of transforming distributions. In our proposed method, losses are extracted at two levels as shown in (figure 3.4). The first loss is content loss and it is extracted at end of the generator and is used to train generator. It is calculated on generator's output directly by comparing outputs of first convolutions of the generator.

The second loss is wasserstein loss and is calculated towards the end on outputs of the whole model. The discriminator is trained on wasserstein loss. Representing different components of model mathematically:

Discriminator with its components: $D_{øD}$
Generator with its component: $G_{øG}$
Blur image: I^B
Sharp image: I^S
Dimensions of feature map: $W_{i,j}$, $H_{i,j}$

Feature maps after j-th convolution and before i-th max pooling layer: $\Theta_{i,j}$

ʎ= 100

Total loss is sum of WGAN loss and content loss:

$$L_{ToTAL} = L_{GAN} + \lambda L_C \quad \text{..................3.2}$$

$$L_{GAN} = \sum_{n=1}^{N} -D_{\theta D}(G_{\theta G}(I^B)) \quad \text{....................3.3}$$

$$L_C = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^S)_{x,y} - \phi_{i,j}(G_{\theta G}(I^B)_{x,y})^2 \quad \text{....................3.4}$$

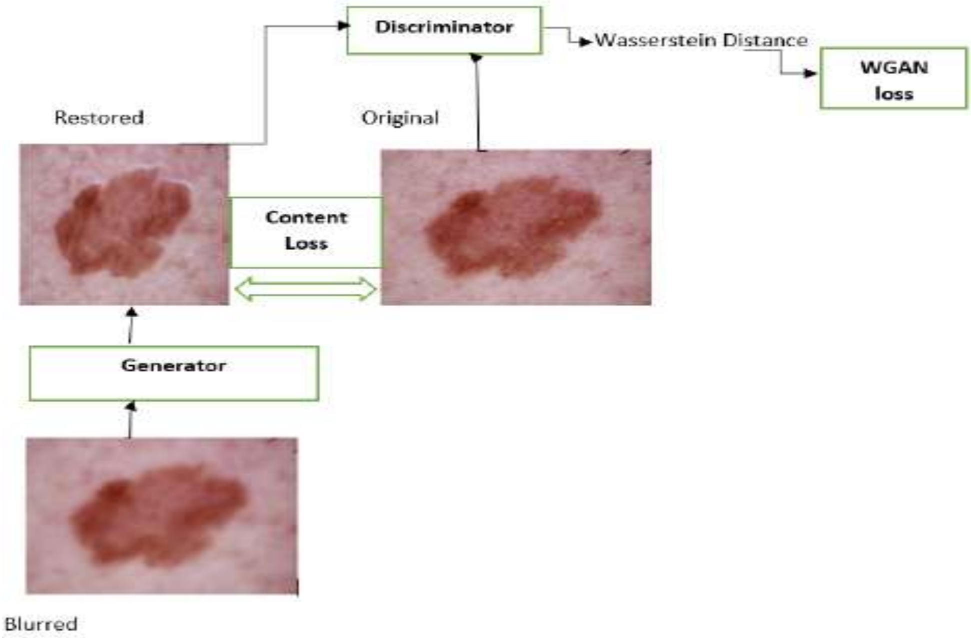

Fig. 8 SRGAN training (GitHub)

D. THE TRAINING

The training in SRGAN occurs in an alternating way similar to the original GAN. Unlike in the original GAN where the generator didn't have direct feedback and needed to depend entirely on discriminator's loss, SRGAN has direct feedback on the generator's output. It is because generator takes in real blurred images as input than random noise. In the first step, the generator generates fake images.

The discriminator is trained by taking in both original images and restored images generated by the generator to compute wasserstein distance between them. After the first step, we have a good approximator for the wasserstein distance. Then, in next step the research work optimize the generator to reduce this distance and each time, the discriminator fails to distinguish images generated by the generator, penalization term is added to gradient of discriminator and it updates its weights. Here, the training is done in an alternating manner. It follows the following steps:

- ➢ Load the data and initialize all the models
- ➢ Add Adam optimizer
- ➢ Set trainable option in Keras to prevent discriminator from training

- Start launching epochs
- Divide dataset into batches
- Train discriminator and generator successively on basis of both losses
- Generate fake inputs with generator
- Train discriminator to distinguish fake from real inputs
- Train whole model

In the Leaky ReLU (it is an activation function similar to ReLU but it provides a small and positive gradient when unit is not in active state), the slope of the leak was set to leak = 0.2. the research work followed training approach in [30], and applied gradient descent optimization on both discriminator and generator with Adam as a solver incorporating first and second moments of gradients, controlled by beta1 = 0.9 and beta2 = 0.999 respectively. We used learning rate of 0.0001 for 150 epochs.

V. THE PROPOSED ALGORITHM

The proposed method involves two separate steps: constructing an ensemble of CNN and training it with an enlarged dataset. Here, an ensemble is constructed by manipulating the training example in order to generate numerous hypotheses. Algorithm is trained several times, each time on a different subset of training samples.

Such construction method is suitable for CNN due to its unstable nature. Typical traditional ensemble methods like random forest assigns more than 30 decision trees, but in case of CNN, due to training constraints like computational expenses, time consumption, etc. The selection of five (5) to ten (10) CNN networks for an ensemble has been considered. Here, we selected five (5) CNN networks for an ensemble. We trained multiple networks and the final classification is obtained by taking the average of the probabilities returned by each network for each class label. Mathematically, final classifier decision is represented as:

$$\sum_{n=i} w_i h_i(x)$$

........................(4.5)

Where in each iteration i, the weighted error are calculated and weights are updated and the hypothesis is returned. Here, each classifier is a hypothesis about true function f. In other words, hi represents the hypothesis that a classifier returns true function f when it faces any new case 'x' and hypotheses space in an ensemble represents a space of multiple hypothesis. Weights are updated till minimum error is observed and the best hypothesis is returned. Working of the averaging prediction over multiple networks is explained by Jensen's inequality in the section 4.2.1. We train an ensemble of five CNNs on enlarged data set comprising original samples, traditionally augmented samples and GAN generated samples. With enough training data, the ensemble explores space of all possible classifiers. Therefore, in this chapter, we propose a fusion of ensemble of CNNs with augmentation techniques as shown in figure below.

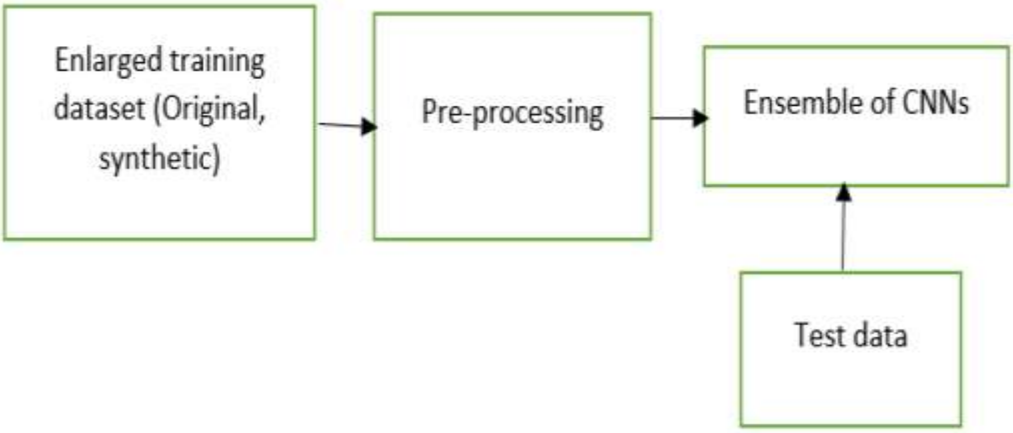

Fig. 9 Flowchart of proposed method

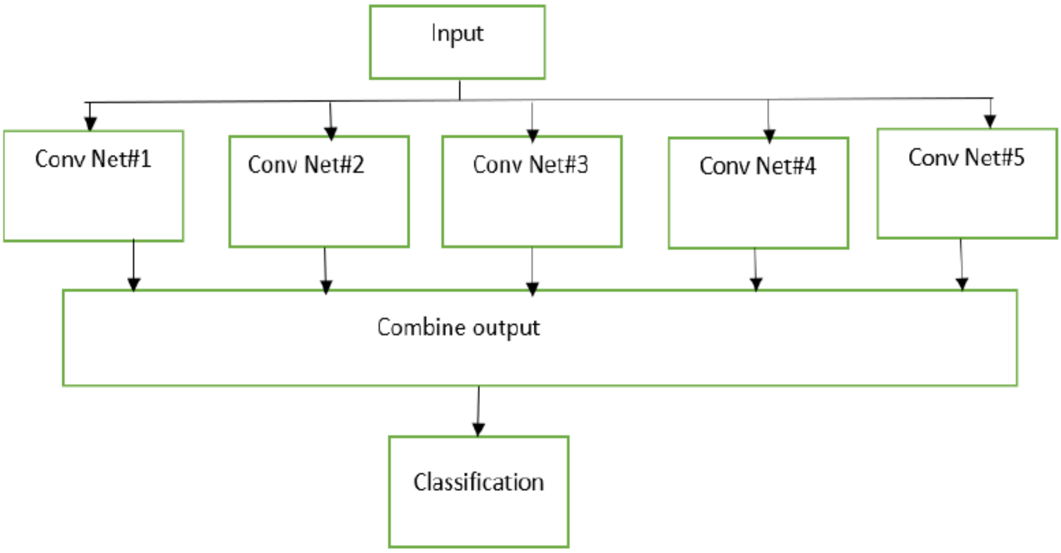

Fig. 10 Flowchart of Ensemble of CNNs

A. TRAINING AN ENSEMBLE OF CNN

Multiple networks are trained simultaneously using a for-loop to train five networks in our case. This output in a serialized model at the end of each iteration.

First, five different CNN models are constructed using proposed CNN architecture as mentioned in (chapter 3). Then, models are loaded and trained on enlarged dataset of skin lesions using stochastic gradient optimizer and binary cross entropy as cost function. Stochastic gradient descent optimizer is initialized using learning rate alpha = 0.001, momentum = 0.9, number of epochs = 40, decay = 0.001/40, and used Nesterov accelerations. The networks are trained for 40 epochs using batch size of 64. Since, five CNN models are trained consecutively, training consumes 5 times more time when compared to training a single CNN model. An ensemble is evaluated on the test dataset where each model in an ensemble produces 2 probabilities for each class label for every data point in the test dataset i.e each model produces prediction array of size 100X2 (test dataset has 100 images). Then, predictions from all models are accumulated by looping over each individual model.

After looping over five models, new prediction array has a representation of (5, 100, 2) representing five models with 2 class label probabilities for 100 test data points. Then, the average of probabilities for each test data point across all five models are taken, which is the final output of ensemble method.

The training of an ensemble of CNN follows the same steps as training an individual CNN classifier in (chapter 3). The input to an ensemble of CNNs in our work consists of 6000 training data data (number of original samples = 1000, number of traditionally augmented samples = 4000, SRGAN generated samples = 1000). It follows the same preprocessing techniques and each individual CNN classifiers are trained in similar way. The only difference lies in the final portion, where the outputs of each classifier are put together into a single meta-classifier.

B. THE EXPERIMENTAL SETUP

Dataset used in this work consists of coloureddermoscopic images collected from the International Skin Imaging Collaboration (ISIC), 2018 archive. Data were presented in the form of challenge ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection by [33], [34]. Since the research work is focused on small datasets, we selected 1000 images of benign skin lesion and 1000 images of malignant skin lesion from archive. Benign keratosis skin lesion was selected as benign and melanoma as malignant. Further, 100 benign and 100 malignant images are collected for the work test set from the archive. Out of different available data sources, ISIC images are in standard format and have been processed to facilitate researchers working towards melanoma detection. All training images were resized to 64 by 64 pixel sizes before feeding into pre-processing pipeline and classifier models.

The simulation was performed on a NVIDIA Quadro P4000 GPU using Python (Keras library). Total simulation included three major segments: training the CNN model, training the

SRGAN generator and the discriminator model, and training an ensemble of CNNs. Incorporating the traditionally augmented images, a CNN model was trained to get the baseline result referred as CNN-AUG (model I). Then, the SRGAN generator and discriminator were trained to get the synthetic images.

Again, incorporating synthetic images, a CNN model was trained to get our second result which is referred in this section as CNN-AUG-GAN (model II). Finally, an ensemble of CNNs model was trained on the final dataset to get the final result which is referred to in this section as EN-CNN-AUG-GAN (model III). In the simulation, overall algorithms were executed in the following two stages:

Stage 1: Augmentation: First, skin lesion images were augmented using traditional augmentation techniques as discussed in chapter 3 to augment 4000 images. This resulted in an increase in the dataset size by five times. Second, SRGAN based augmentation produced visually realistic synthetic images. It provided an additional 1000 images. Total enlarged dataset is now comprised of 6000 images. In total, dataset size was increased by six times.

Training Samples	**Number of Images**
Original samples	1000
Traditional augmented + original	1000 + 4000
SRGAN + Traditional + original	1000 + 4000 + 1000
Total images	6000

Table. 1 Augmented Dataset

Stage 2: Classification: Two different types of classification approach were used. The first was classification using single CNN classifier. This resulted in model I and II.

The second approach was an ensemble of multiple CNN classifiers which resulted in simulation of model III.

In this thesis, the CNN classifier is designed as a binary classifier. Accuracy (AC), Sensitivity (SN) and Specificity (SP) are most commonly used metrics to measure performance of classifier. Here, in binary classification, datasets are divided into two classes. SP indicates how CNN classifier predicts a negative class, SN indicates how CNN classifier predicts positive class and AC indicates CNN classifier prediction for both positive and negative classes.

Block diagram of the above stages of augmentation and classification algorithms is shown in (figure 4.1). It consists of six major blocks: Preprocessing, Traditional augmentation, SRGAN based augmentation, CNN classifier, Ensemble of CNN and comparison.

C. PERFORMANCE METRICS OF A CNN CLASSIFIER

To evaluate the performance of classifiers for melanoma detection in this proposed models, Accuracy (AC), Sensitivity (SN), and Specificity (SP) have been used as performance metrics. Performance metrics that are used to evaluate the quality of the classifier were AC, SN and SP. The metrics were calculated from the confusion matrix as shown in (figure 11). For binary classification, the classifier takes all test images and predicts either positive or negative with four cases explained below:

1. True Positive (TP): When prediction is correct and positive then it is called true positive.

Example: Skin lesion is detected as benign (positive class) when it is benign (positive class).

2. False Positive (FP): When prediction is incorrect but positive it is called false positive.

Example: Skin lesion is detected as benign (positive class) when it is melanoma (negative class).

3. True Negative (TN): When prediction is correct but negative then it is called true negative.

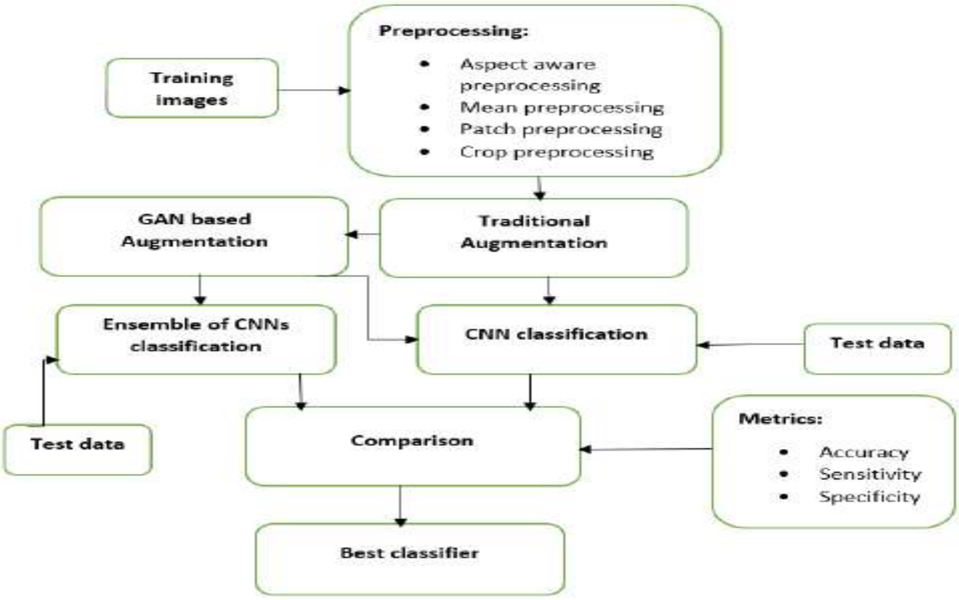

Fig. 11 Proposed skin lesion classification methodology

Example: Skin lesion is detected as melanoma (negative class) when it is melanoma (negative class).

4. False Negative (FN): When prediction is incorrect and negative then it is called false negative.

Example: Skin lesion is detected as melanoma (negative class) when it is benign (positive class).

Above information is illustrated in confusion matrix in (figure 12). Here, 0 indicates a positive class and 1 indicates a negative class. The above information can be depicted in the following confusion matrix for a binary classification problem as shown in the (figure 12).

		PREDICTED LABEL	
		0	1
TRUE LABEL	0	TP	FN
	1	FP	TN

Fig 12 Confusion Matrix

Based on the four possible outcomes, performance metrics AC, SN and SP are defined below:

1. Accuracy (AC): Accuracy is defined as ratio of total numbers of correct predictions to total number of images in dataset. Total numbers of

correct predictions is calculated as sum of TN and TP. In formula, AC is expressed as:

$$AC = \frac{TP + TN}{TN + TN + FN + FP} \ldots\ldots\ldots\ldots\text{(4.1)}$$

The best value for accuracy is 1 and the worst is 0.

2. Sensitivity (SN): Sensitivity is defined as ratio of true positive predictions to a total number of positive predictions. Total number of positive prediction is the sum of TP and FN, so the formula for SN can be expressed as:

$$SN = \frac{TP}{TP + FN} \ldots\ldots\ldots\ldots\ldots\ldots\text{(4.2)}$$

The best value for sensitivity is 1 and the worst is 0.

3. Specificity (SP): Specificity is defined as ratio of total number of true negative predictions to a total number of negative predictions. Total number of negative prediction is sum of TN and FP, so the formula for SP can be expressed as:

$$SP = \frac{TN}{TN + FP} \ldots\ldots\ldots\ldots\text{ (4.3)}$$

The best value for specificity is 1 and the worst is 0.

D. SIMULATION RESULTS

The research baseline model was model I (CNN-AUG). Initially, using the traditional augmentation, 4000 images for each class were generated which was optimal point for traditional augmentation. Then, effects of traditional data augmentation for skin lesion classification have been examined. Classification performance by model I yielded 77% sensitivity, 81.38% specificity and 79.29% of accuracy. In model II (CNN-AUG-GAN), skin lesion was synthesized using SRGAN. By adding synthetic data augmentation, performance of CNN classifier increased to 80.26% sensitivity, 82.78% specificity and 81.32% of accuracy. Then, the single CNN classifier was replaced by an ensemble of five CNNs in model II and this fusion of ensemble of CNN with GAN based augmentation resulted in model III (EN-CNN-AUG-GAN). Model III yielded 84.32% sensitivity, 84.21% specificity and 84.30% accuracy. Higher value of accuracy is not solely a good indicator for any classification model with smaller training samples. Therefore, to confirm that model III is a better performing classifier compared to model I and II, then sensitivity and specificity have been evaluated. It is evident from result that model III performed well in terms of all metrics. It can classify TN (melanoma) class, TP (benign) class and both TN and TP by around 84% of accuracy. Better result in specificity and sensitivity confirms high accuracy and implies that an ensemble of CNN classifiers is better at separating both positive and negative classes compared to other model. Table 2 provides overall view of these values.

Model No.	Model	Sensitivity	Specificity	Accuracy
1	CNN – AUG	0.7702	0.8138	0.7929
2	CNN – AUG – GAN	0.8026	0.8278	0.8132
3	EN – CNN – AUG – GAN	0.8432	0.8421	0.8430

Table:2 Accuracy, Sensitivity and Specificity of CNN classifier for melanoma detection.

Graph in (figure 12) illustrates the result of three models in terms of dataset size. Baseline result is denoted by dotted lines. With no augmentation, the performance of CNN was around 63% due to overfitting issue faced by smaller dataset. Adding augmented images by traditional augmentation improved performance of CNN and also, removed overfitting issues and acted as regularizer. Accuracy was improved till 79% and then saturated. After this point, adding more data by traditional augmentation has no effect in improving accuracy. Smooth lines in the graph shows total accuracy of CNN classifier after addition of synthetic data generated by SRGAN. Classification result improved from 79% to 81% for 6000 samples per class. After 6000 samples per class, adding more synthetic data has no effect in improving accuracy due to less diversity for this particular dataset. By keeping samples per class constant and instead of using single CNN, ensemble of five CNNs into a meta-classifier resulted in improvement in accuracy to 84%. When compared to baseline model, performance improved by 5% in accuracy, 7% in sensitivity, 3% in specificity. Improvement in sensitivity reduces chances of patient with melanoma being classified as benign, which is crucial in case of melanoma detection.

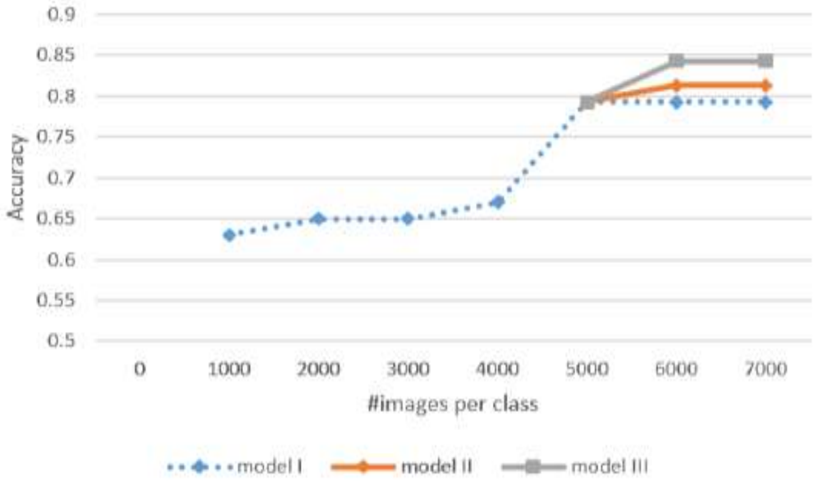

Fig. 12 Total accuracy for skin lesion classification of melanoma and benign skin lesions by increasing training data size.

E. VISUAL RESULTS OF AN AUGMENTATION TECHNIQUE

The performance of a CNN classifier for the skin lesion classification was confirmed by the quantitative values provided in section 4.2. Result in section 4.2 and comparative analysis in discussion section confirmed the role of the GAN based augmentation in the classification. However, the visual quality of synthetic images generated by GAN can also support the importance of GAN. Here, SRGAN, a version of CGAN, produced visually realistic images of skin lesions via conditional learning. The images generated by SRGAN are shown below in figure 13.The images generated by this method overcame artifacts problem prevalent in other version of GANs.

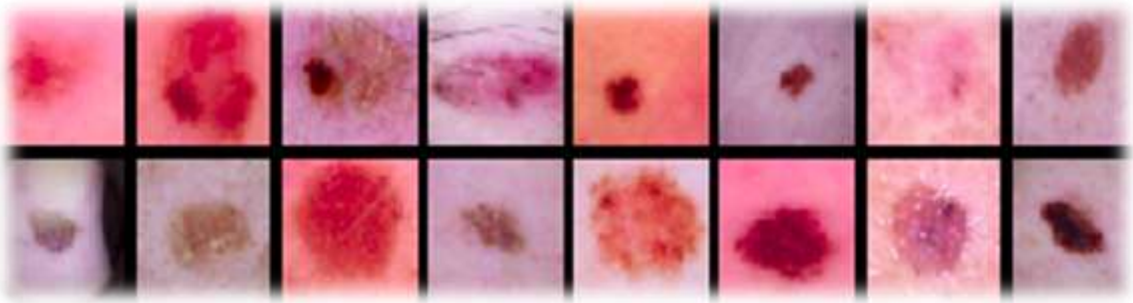

Fig. 13 Synthetic images generated by SRGAN

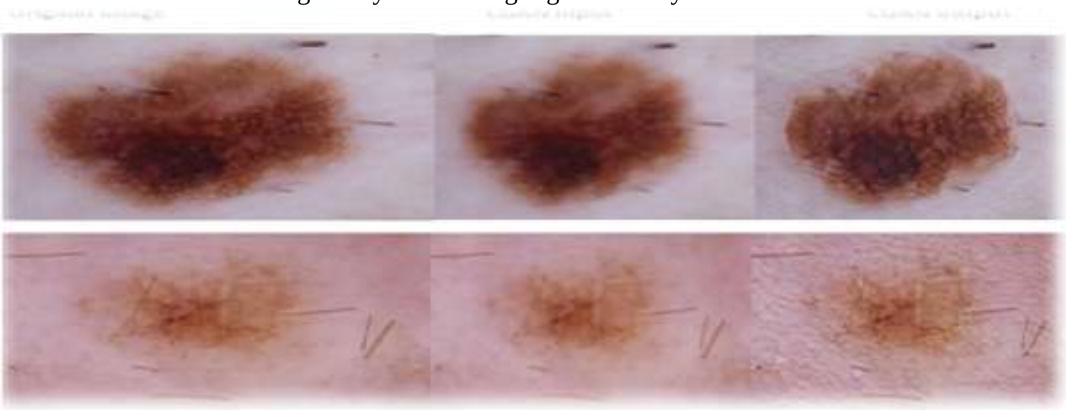

Fig. 14 Synthetic images generated by SRGAN (Malignant (Top) andBenign (Down)

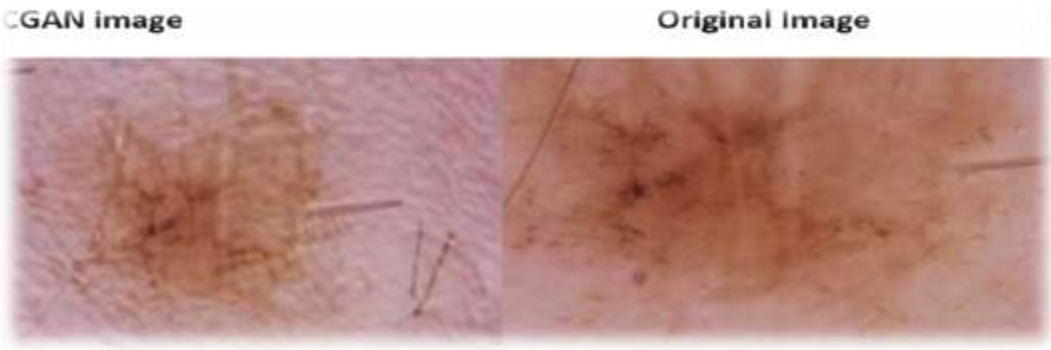

Fig. 15 Zoomed version of SRGAN image (left) and original image(right).
Here original image is zoomed in deeper than SRGAN image to show resolution of original image.

F. SUMMARY AND FINDINGS

This research work presents a novel classification system to address data scarcity issues in deep learning using ensemble learning and different augmentation techniques. A state of- the-art deep learning classifier system for melanoma detection was built, where two classes melanoma and benign were taken as training samples. Initially, the number of training samples was 1000 per class. To train a deep learning system, generally, 4000 or more samples per class are required. Therefore, this thesis proposed three different models and each model involved two main stages: augmentation and classification.

For augmentation, this thesis proposed traditional augmentation and neural network based augmentation (GAN). SRGAN produced visually realistic synthetic images, which added diversity to dataset. From augmentation, datasets were enlarged by six times and were enough to train convolutional neural networks sufficiently by overcoming overfitting issues as well as improving performance of CNN classifier in terms of accuracy, sensitivity and specificity. For classification, this thesis proposed training using a single CNN classifier and an ensemble of CNN classifiers. Ensemble method combined five CNN classifiers into a single-meta classifier that produced result based on Jensen's inequality. Model I implemented traditional augmentation to train single CNN classifier. Model II proposed implementation of SRGAN based augmentation to train single CNN classifier. Model III proposed implementation of enlarged dataset by traditional augmentation and SRGAN based augmentation to train ensemble of CNNs meta-classifier. Model II performed better than model I indicating GAN based augmentation as suitable approach to enlarge dataset size. Similarly, Model III performed best amongst all proposed models, therefore, use of ensemble approach in CNN along with augmentation techniques appear to be more effective for classification problem dealing with data scarcity issues.

After implementing three models, the following are best findings that can be referred to while dealing with a smaller training dataset:

1. Traditional augmentation can prevent overfitting and can improve generalization in CNN classifier.
2. After optimal point is achieved in traditional augmentation, it is better to use SRGAN based augmentation to increase dataset size. It can prevent overfitting and also, improves performance of CNN classifier.
3. Ensemble of CNNs can improve performance of CNN classification by 1 to 5%. Implementing an ensemble of CNNs with augmentation techniques mentioned above can improve overall performance to a great extent.

This research has discussed various neural network techniques for skin cancer detection and classification. All of these techniques are noninvasive. Skin cancer detection requires multiple stages, such as preprocessing and image segmentation, followed by feature extraction and classification. Each algorithm has its advantages and disadvantages. Proper selection of the classification technique is the core point for best results. However, CNN gives better results thanothers.

G. FUTURE WORK

The creation of large public image datasets with images as representative of the world's people as possible to avoid racial bias is another major task in this research field. Image prejudice based on gender and race AI prejudice means that the models and algorithms fail to give optimal results for people of an under-represented gender or ethnicity.

Mostly, skin lesions from light-coloured skin can be seen in current datasets. For example, the ISIC and GitHub datasets images are mostly obtained in the USA, Europe and Australia.

In addition, CNNs try to extract the skin colour to achieve a proper classification for dark skinned humans. This can only happen if the training dataset contains sufficient images of dark-skinned people. The size of the lesion also has significant importance. If the lesion size is smaller than 6mm, melanoma cannot be detected easily.

The addition of clinical data such as race, age, gender, skin type, as inputs for classifiers may also help to increase classification accuracy.

VI. CONCLUTION

There are two major problems with skin lesion classification using CNN. One is insufficient data and the other is class imbalance and to address both of the problems we need more data. Traditional augmentation can help us up to a certain extent after that we have to search for novel methods. After observing progress made by GAN in medical image augmentation, we conclude that GAN is one of the ways which we can surely increase the quality and quantity of dataset. This will help us to increase the performance of the classification of skin lesions.