Smart Intelligent Computing and Communication Technology
V.D. Ambeth Kumar et al. (Eds.)

doi:10.3233/APC210043

Data Set Preparation Using the Method of Data Augmentation for Classification of Skin Diseases

AdlinLayola J.A[a,1] and V. MuraliBhaskharan[b]
[a]*PG Student, Dept of CSE, Rajalakshmi Engineering College, Chennai, India*
[b]*Professor, Dept of CSE, Rajalakshmi Engineering College, Chennai, India*

Abstract. Skin disease are mostly ignored and provided less importance at theearly stages. Some ignorance among people might lead to skin cancer. In existing approach, the increased skin disease are identified at the later stage using biopsy only. The inspection is performed manually by considering many histopathological features.This paper explains the development of a model which helps to detect the various skin diseases at early stage using neural networks.For classification and prediction purpose this paper using Convolutional Neural Network.For improving the accuracy of prediction we are going to do dataset preparation. Data augmentation is using for increasing the count size of the training data. So that the accuracy of the classification will increase and will get 97% result.

Keywords. Skin Disease, Dataset Preprocessing, Data Augmentation, Deep Learning, CNN.

1. Introduction

Skin diseases are most common and difficult diseases for diagnosis because of its lack of awareness and ignorance. In many developing countries also people consult dermatologist for skin disease and prevention measures. Importance of skin disease without ignoring at the early stage is very important as skin plays a major role in protecting the human body against fungal and harmful bacterial infections. Many people get skin disease through their inheritance, job, lack of nutrition, regular habitats, exposed to chemicals etc. season, winter season. Thus identifying skin disease and diagnosis at the early stage is very crucial.To overcome this problem an application is identified and is used to predict the skin disease by using convolutional neural network. In this research, a classifier is created and which will produce different class for each skin disease.

[1] AdlinLayola J.A, PG Student, Dept of CSE, Rajalakshmi Engineering College, Chennai, India;
E:mail:adlinlayola @gmail.com

The classes are generated by analyzing the input image and the image is compared matching with the previous training data. By this process, can improve the accuracy of prediction. This research is followed by HAM10000 dataset[1] which has approximately 10015 images of various skin diseases. The dataset is used by both testing and training phase. The dataset should be prepared for improving the accuracy. Increasing the training sample will give more accuracy so that the count size of the dataset is increased by data augmentation. The classifier using the augmented data set for prediction. Thus the result will show high accuracy. The following parts of this paper are arranged as follows: section 2 is a literature review, section 3 is the proposed method, section 4 is the results and discussion, and section 5 is the conclusion.

2. Literature survey

A review of existing research gives certain idea about this work. In [2] this paper describe the advantages of data augmentation. This paper explains the performance of classification improvement using data augmentation. In [3] proposed a approach for improving the age estimation using Convolutional Neural Network. Here the approach is evaluated for producing more accuracy by using data augmentation method. The results shows that the results were good when the classifier was trained with data augmentation. In [4] the training data set is expanded. Even after expanding the data set also the efficiency of deep convolutional neural networks will be improved. For experiment results they using PCA jittering, Noise,AN/WGAN, Rotation, Shifting, Flipping, Color jittering, Cropping data augmentation operations. This paper mainly discussed about the comparision of accuracy results for augmented and un augmented dataset. In [5] The additional samples are given by mapping elements from a different pool rather than from the dataset itself, according to a new data augmentation approach proposed. Cross-Dataset Data Augmentation was proposed and demonstrated sucessfully. In [6] compared and analyzed multiple operations of data augmentation. And new operation has been introduced which means image style transfer. In [7] evaluating the data augmentation in Charcoal Image Classification. This paper explores sub-images and morphological transformation as data augmentation approaches and produced 99.36% as average accuracy. In [8] exploits data augmentation approach in brain tumor examples and shows the boost up of generalization abilities of deep learning. In this approach, data augmentation techniques are applied to magnetic resources images and the advances were reviewed. In [9] focuses two data augmentation techniques which are oversampling and data warping. This survey gives different solutions for reducing the overfitting problem. In [10] build a deep learning study on skin disease image recognition.The review's finding is that data augmentation increases classification accuracy[11-13].

3. Proposed Methodologies

Methodology Our proposed system incorporates two technologies, they are dataset preparation and deep learning algorithm in which this model using convolutional neural network. Figure 1 shows the flow of classification of skin diseases followed by data

augmentation.The dataset was created using a dermatoscopic image of common pigmented skin diseases from Harvard's HAM (Human against Machine) dataset.Because of the unequal and scattered number of images in each class, we improved the dataset with the aid of data augmentation using Keras.Because of the difference in the number of images, we decided to build a dataset to improve the dataset's consistency, which will improve the model's accuracy.A large dataset is required for the CNN model to perform well, so high performance can be achieved by augmenting the available data.Data augmentation increases the size of the dataset, which improves the model's accuracy. Rotation, shearing, zooming, cropping, rotating, and adjusting the brightness level are examples of data augmentation operations. After augmentation the classification is done using convolutional neural network.

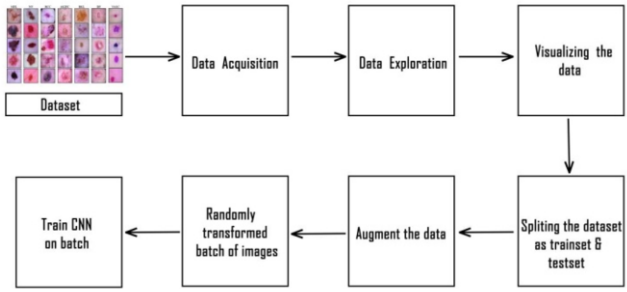

Figure 1. Flow diagram of skin disease classification followed by data augmentation

3.1 Dataset Collection

Skin disease dataset are collecting various images as a sample pictures. This paper has collected many skin infected image from various parts of the human body. In order to have more pictures of skin diseases, this paper is using HAM 10000 dataset which has seven different skin diseases which are vascular lesion(vas), benign keratosis lesions(bkl), melanoma(mel), actinic keratoses(akiec), melanocytic nevi(nv), dermatofibroma(df) and basal cell carcinoma(bcc).These are used for trainingand testing. Figure 2 displays different skin disease sample images of HAM10000 dataset.

Figure 2. Different skin disease sample images of HAM10000 dataset

3.2 Data Preparation

Keras Preprocessing means data preprocessing alongside the info augmentation module of the Keras deep learning library. This process includes the work along with image data, text data, and sequence data. Keras Preprocessing is compatible with Python 3.6. initially clearing the dataset of null values takes place. one-hot encoding is used in this process to convert categorical variables to numerical variables. Preprocessing technique is utilized before the deep learning algorithm, where in that all raw imagesare transformed and given to the classification. The training of the raw image on the convolutional neural network, leads poor performance of classification.

3.3 Augmentation of Data

The dataset which has been selected are made to be enlarged. If the process of increasing the training samples will result in more accuracy. HAM 10000 dataset consists of seven different classes of skin diseases. Among this only one class has maximum number of the sample images. According to that maximum number this paper explains, the method of augmenting the other classes of skin disease. As a result all the classes approximately consist of equal number of samples. For augmentation Brightness range, rotation range, zoom range, width shift range, horizontal flip, vertical flip, height shift range operations are used.

3.4 Classification and Prediction using CNN

Initially classifier has been created. Suppose an input image is given, preprocessing along with feature extraction takes place. Theseprocess are carried out in different layers. In the final layer it combines the extracted feature and represent it in the new model. Thus training the model, will predict the skin disease with high accuracy. Finally this paper ensures the accuracy with 97% approximately.

Results and discussion are described in the later stage. The model along with experimental approach is provided.

4. Results And Discussion

This paper explains the details about the classification of skin diseases followed by data set preparation using data augmentation technique.

4.1 Data Exploration

Data exploration is the process of visualizing the data from the dataset. Here we can see how many images are available in each classes with count sizes. Figure 3 explains the actual count size of the images in eachclasses.

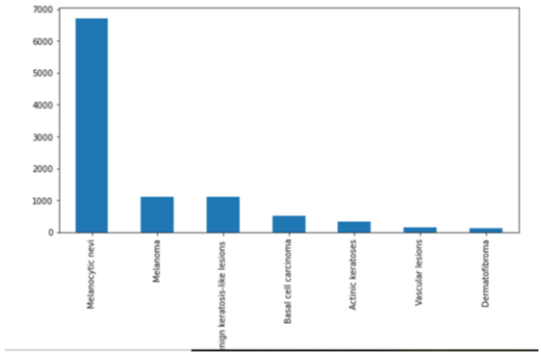

Figure 3. Actual count of images in each classes

4.2 Data Augmentation

For improving the accuracy of the classification have to increase the image count size of training sets. For this initially have to import the libraries. After that creating new directory for the images. Inside that base directory another new directories are created for training and testing of each classes. From the meta data have to divide the images as testing set or validation set and training set. Transfer the images to training and testing sets. Then augment the data using image data generator tool. Figure 4 shows the count size of training dataset after augmentation

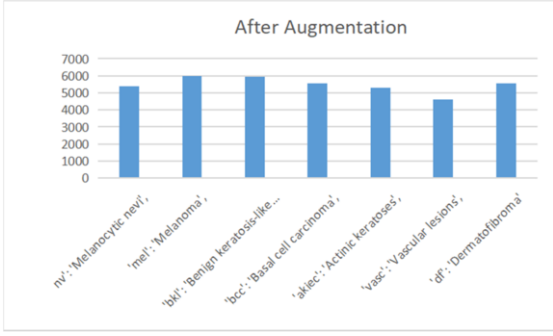

Figure 4. Increased count size of images in each classes after augmentation

We can also visualize the augmented images from the training dataset directory. Figure 5 shows some sample images of skin lesions which are augmented. Here, rotation, zoom, horizontal flip, vertical flip, height shift range, width shift range, brightness augmented operations are used.

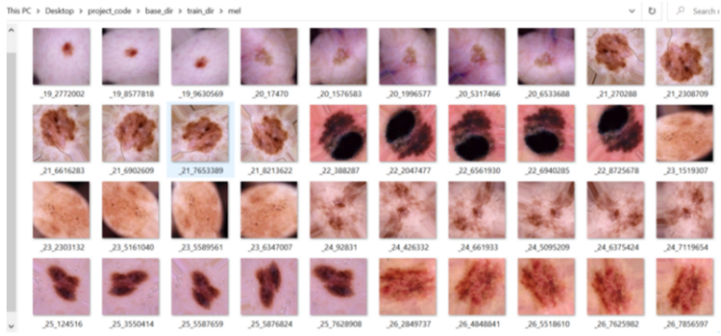

Figure 5. Sample images of skin lesions after augmentation

4.3 Classification

Finally classifier is created and the skin disease is predicted. Here using convolutional Neural Network is used for classification. And finally getting 99% accuracy results for predicting the skin disease by using data augmentation. Figure 6 represents the skin disease classification and prediction using the Convolutional Neural Network.

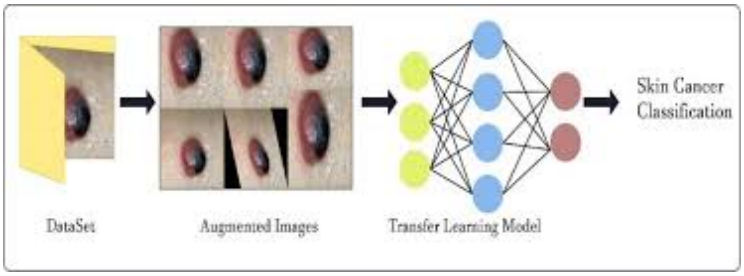

Figure 6. Classification process of skin diseases after data augmentation

5. Conclusion

After study and implementation of deep learning in the field of skin diseases it is concluded that Convolutional Neural Network while using data augmentation method provide very competitive results comparing to the state-of-the-art. The network is trained for the test of different skin diseases. After training the network was able to classify the skin cancer images into different classes. Forthis current work 99% accuracy was achieved with a CNN usage of data augmentation. Compared to previous works, this work uses more number of images in order to achieve better performance. Once improvements are made on the Classification deep learning algorithms for the skin disease in the field of medical the results can be improved. So nowadays deep learning is the best solution of skin diseases classification and recognition of cancerous diseases.

References

[1] Philipp Tschandl1 , Cliff Rosendahl2 &Harald Kittler(2018), The HAM10000 dataset, a large collection of multi-source dermatoscopicimagesof common pigmented skin lesions, www.nature,com/scientific data, 5:180161.

[2] Sebastien C. Wong,Adam Gatt, Victor Stamatescu and Mark D. McDonnell (2016), Understanding data augmentation for classifification: when to warp?, Institute of Electrical and Electronics Engineers,

[3] Italo de Pontes Oliveira, Joao Lucas Peixoto Medeiros, VinciusFernandes de Sousa Adalberto Gomes Teixeira Junior, Eanes Torres Pereira, Herman Martins Gomes (2016), A Data Augmentation Methodology to Improve Age Estimation using Convolutional Neural Networks, 29th SIBGRAPI Conference on Graphics, Patterns and Images,2377-5416/16.

[4] Shijie.J, W. Ping, J. Peiyi and H. Siping(2017), Research on data augmentation for image classification based on convolution neural networks, 2017 Chinese Automation Congress (CAC), Jinan, China, pp. 4165-4170.

[5] Gasparetto.A et al.(2018), Cross-Dataset Data Augmentation for Convolutional Neural Networks Training, 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, pp. 910-915.

[6] Mikołajczyk.A and M. Grochowski(2018), Data augmentation for improving deep learning in image classification problem, 2018International Interdisciplinary PhD Workshop (IIPhDW), Świnoujście, Poland, pp. 117-122.

[7] L. T. Menon, I. A. Laurensi, M. C. Penna, L. E. S. Oliveira and A. S. Britto(2019), Data Augmentation and Transfer Learning Applied to Charcoal Image Classification, 2019 International Conference on Systems, Signals and Image Processing (IWSSIP), Osijek, Croatia, pp. 69-74.

[8] Nalepa, Jakub et al (2019), Data Augmentation for Brain-Tumor Segmentation: A Review, Frontiers in computational neuroscience vol. 13 83.

[9] C. Khosla and B. S. Saini (2020). Enhancing Performance of Deep Learning Models with different Data Augmentation Techniques: A Survey, 2020 International Conference on Intelligent Engineering and Management (ICIEM), London, UK, pp. 79-85.

[10] Ling-Fang Li, Xu Wang, Wei-Jian Hu , Neal n. Xiong , (Senior Member, Ieee), Yong-Xing Du, And Bao-Shan Li1(2020).Deep Learning in Skin Disease Image Recognition: A Review. IEEE access, volume 8.

[11] Ambeth Kumar V.D, Dr.M.Ramakrishnan, V.D.Ashok Kumar and Dr.S.Malathi (2015).Performance Improvement using an Automation System for Recognition of Multiple Parametric Features based on Human Footprint. kuwait journal of science , Vol 42, No 1 (2015), pp:109-132.

[12] Ambeth Kumar.V.D, V.D.Ashok Kumar, S.Malathi, K.Vengatesan, M.Ramakrishnan .Facial Recognition System For Suspect Identification Using A Surveillance Camera . Pattern Recognition And Image Analysis (Springer), Volume 28, Issue 3, Pp 410–420, 2018. (DOI: 10.1134/S1054661818030136)

[13] Ambeth Kumar.V.D and M.Ramakrishan .Employment Of Footprint Recognition System. In The Month Of December For Indian Journal Of Computer Science And Engineering (IJCSE) Vol. 3 No.6 Dec 2013.