Novel Dataset Generation for Indian Brinjal Plant Using Image Data Augmentation

Balwant Gorad¹, Dr. Sirbi Kotrappa²

¹Research Scholar, Computer Science and Engineering, KLE Dr. MSSCET, Belagavi, Karnataka, India – 590 008
²Professor, Computer Science and Engineering, KLE Dr. MSSCET, Belagavi, Karnataka, India – 590 008

goradbj@gmail.com

Abstract. Machine learning and deep learning have performed outstandingly in many computer tasks to tackle various real-world problems like disease prediction on plants etc. But unfortunately, there are many research areas like disease prediction in the agricultural crops where there is a lack of large-good quality real-world datasets. One way to solve such a problem is by using an available dataset from the internet. The problem of using an available dataset from the internet creates lots of issues. Major issues are using the dataset from different geographical locations which are deployed at other location, model overfitting due to small-sized dataset etc. The main purpose and experimentation done in this research paper are presenting different techniques to increase the size of the Indian Brinjal dataset so that deep learning models can be improved. Here data augmentation techniques to enhance the small-sized image dataset using rotation, channel shift, width shift, height shift, shear transform, brightness, scaling, uniform aspect ratio, zoom, horizontal flipping, and vertical flipping methods are used. At last, a huge high-quality training dataset of size 39,010 is generated from 350 sample images taken from the real field and 1356 high-quality images are generated to validate the model using above mentioned data augmentation techniques.

1. Introduction
Deep convolutional neural network have performed outstandingly in many computer tasks like supervised, unsupervised and semi-supervised learning. But unfortunately, many research areas are not having the good quality dataset with huge volume. One of such domain is Indian agriculture. Quality agriculture is based on collecting comprehensive information from both environment and plant. Information from environment like soil moisture, temperature, wind speed, rain fall, and humidity are to be collected and need to be analyzed. Whereas in case of plant, information like plant disease, plant growth, and insect pest is to be analyzed and monitored. Currently many researchers working in the field of machine and deep learning are using datasets which are available on internet, so satisfactory work is not happening in these areas due to limited size of dataset. To build any kind of machine learning or deep learning model, one of the major task is data acquisition; it’s preprocessing and making good quality dataset ready to build classifiers. Many well-known researchers in these areas suggest having a good quality preprocessed dataset before building the model is better. If you not having good quality dataset, we may lead to less accurate models or infeasible solutions. In case of plant dataset available on internet is from different geographical locations and experiment or
deployment of research models are done different geographical locations, which is leads to infeasible way due to environment factors, soil properties, types of diseases and many other factors. This research-based experimentation we have carried out for Indian Brinjal plant. This is going generate good quality preprocessed dataset with almost having size 40,336 images using various techniques like cleaning data, preparing data and manipulating data using image data augmentation and dimensionality reduction. In our research, the actual data set used in this research to generate high quality dataset in huge volume is collected manually from different farms located in western Maharashtra region. Data Augmentation consists of various techniques that enhance the size and quality of training datasets such that better deep learning models can be built using them. Finally, the outcome of this research experimentation is huge high quality Indian Brinjal dataset which can be used for building various kinds of machine or deep learning classifiers.

2. Related Work

[1] The work carried out in this paper has presented various methods for Data Augmentation, like flipping, color space, translation and noise rejection but to generate good quality dataset, only these methods are not enough. Additions to these all methods, few more methods are needed to be considered like shear transform, scaling, zoom, brightness, channel shift.

[2] The work done in this paper is based on scale & shift parameters, new images are generated from fewer images using these parameters. This researcher focused more about scale and shifting parameters only, but again only these parameters are not enough to generate good quality dataset so there is need to work more in this field.

[4] In this paper, multiple solutions to the problem of data augmentation in image classification are explored and compared. Experiments carried out here used GANs to generate images of different styles. Some questions like safety improvement and quality data generations is not yet achieved.

[5] The work carried out here by TensorFlow community is on flowers dataset. They have performed various data augmentation technique in very good way, results are good but as we need more work to be carry out in Indian agriculture using various data augmentation techniques. So still there is scope to do more work in this field.

[6] The work done in this paper is based on molecular study and enhancement of molecular dataset. They have claimed they have received good result in their work where dataset in is form of text representation of molecule. As they have carried out the work in molecular area, still scope is there, to use different techniques of augmentations in agriculture field.

3. Dataset Collection

3.1. Overview
In this research, work has been carried out on Indian Brinjal plant. We have taken training data set in the form of image. Even it is possible to have dataset in video format, in such case video frames are to be convert into multiple images.

3.2. Dataset Collection for data augmentation
Real world dataset has been taken for image data augmentation process. working with real dataset is always better while building machine or deep learning models which may be used by different researcher from same geographical locations. Along with data collection, information of different disease types on Brinjal plant have been collected from farmers and agriculture institutes/universities from the state of Maharashtra and Karnataka, India.

Total 350 images of Brinjal plant are obtained which corresponds to 7 classes. These 7 classes contain 1 healthy and 6 unhealthy classes which includes alternia leaf spot, bacterial wilt, collar rot, damping off, tobacco mosaic virus and cercospora leaf spot.
Figure 1 Real-time dataset collection

Figure 2 Six common types of brinjal diseases (a) healthy (b) alternia leaf spot, (c) bacterial wilt, (d) collar rot, (e) damping off, (f) tobacco mosaic virus and (g) cersocpera leaf spot

3.3. Overview of Samples Collected

| Type of Brinjal disease   | No. of samples used |
|---------------------------|---------------------|
| Alternia leaf spot        | 40                  |
| Bacterial Wilt            | 45                  |
| Collar Rot                | 35                  |
| Damping Off               | 32                  |
| Tobacco Mosaic Virus (TMV)| 55                  |
| Cersocpera leaf spot      | 27                  |
| Healthy samples           | 116                 |
| **Total samples**         | **350**             |

4. Proposed Method

4.1. Overview of data augmentation

This data augmentation is very powerful tool to create more data. It modifies an image and gives back lots of unique new images, based on different operations. Operations includes rotation, width and height shifts, brightness, shear transform, zoom, channel shift, horizontal and vertical flips, data normalization, rescale and pre-processing function. These all operations plays a vital role to generate
big volume good quality dataset with little care taken during the process. The key idea of all these methods is explained in following part of this proposed method section

4.2. Uniform Aspect Ratio
First step is to ensure that, all the images present are dataset are having the same size and same aspect ratio.

4.3. Image Scaling
Once it has been ensured of have same dimensions to all images (or have some predetermined aspect ratio), it is the time to perform scaling operation on each image appropriately. We have decided to have images with width of 400 and height of 300 pixels.

4.4. Rotation
This method is used to rotate image by specified degree. If we set rotation range as $60^0$ then rotation of image is done between $-60^0$ to $+60^0$. You can experiment part to know more about this rotation.

4.5. Width and Height Shifts
This method controls the amount of width and height shifting. It is number of floating type and its value is in between 0 and 1. This specifies upper bound by which image is randomly shifted on left or right side. In case of image is randomly shifted towards up or down.

4.6. Brightness
The brightness range indicates the range for randomly picked image brightness shift value. 0.0 indicates absolutely there is no brightness, and 1.0 indicates maximum image brightness.

4.7. Shear Transform
Shear transformation change the shape of the given image. Shear transforms and rotation is different from each other. In shear transformation, we are fixing one axis and stretching the image at a certain angle called shear angle. This is going to create a sort of ‘stretch’ in the given image, which we are not able to see rotation. Shear range specifies the angle of the slant in degrees.

4.8. Zoom
This random zoom technique is obtained by the zoom range argument in Keras with TensorFlow machine learning framework. A zoom with less than 1.0 magnifies the image, while a zoom with greater than 1.0 zooms out of the image.

4.9. Channel Shift
This Channel shift technique randomly shifts the channel values by using random value which is chosen from the range specified by channel shift range.

4.10. Horizontal and Vertical Flips
The horizontal flipping technique generate horizontally flipped images on a random basis whereas generates vertically flipped images.

5. Experiments

5.1. Overview of data augmentation
These experiments have been carried out using Google TensorFlow with Keras. Dataset required for these experiments synchronized with Google Drive and it is then mounted with Google collaboratory. The purpose of conducting these experiments on collaboratory is to have high end hardware configuration. During the experimentation 12.72GB RAM was available and 107.77GB disk was available.

5.2. Uniform Aspect Ratio
This technique ensures, all the images are of same size and same aspect ratio. When we have captured images from the field, it was having different uniform aspect ratio, so we have converted all images into same uniform aspect ratio.

5.3. Image Scaling
We have decided to have images with width of 400 and height of 300 pixels.

5.4. Rotation
Rotation range $= 60^0$
5.5. Width and Height Shifts

Width shift = 30, Height shift = 10

5.6. Brightness

Brightness range = (0.5, 2.0)

5.7. Shear Transform

Shear range = 40
5.8. Zoom

Zoom range = [0.7, 1.5]

5.9. Channel Shift

Channel shift range = 100

5.10. Horizontal and Vertical Flips

Horizontal flip = True
6. Results

Several experiments are carried out to explore effectiveness of different image data augmentation techniques. From each disease class, sample images are used to augment it. After augmenting huge sized dataset contains 40336 for both training and validation is generated.

Disease type vs. No. of augmented images for training dataset are shown below.

Table 2 Disease wise augmented images for training

| Type of Brinjal disease  | No. of samples used | No. of Augmented Images |
|-------------------------|---------------------|-------------------------|
| Alternia leaf spot      | 40                  | 1907                    |
| Bacterial Wilt          | 45                  | 3136                    |
| Collar Rot              | 35                  | 2621                    |
| Damping Off             | 32                  | 3608                    |
| Tobacco Mosaic Virus (TMV) | 55                  | 3302                    |
| Cercospora leaf spot    | 27                  | 1835                    |
| Healthy samples         | 116                 | 22601                   |
| Total                   | 350                 | 39010                   |

Disease type vs. No. of augmented images for validation dataset are shown below.

Table 3 Disease wise augmented images for validation

| Type of Brinjal disease  | No. of samples used | No. of Augmented Images |
|-------------------------|---------------------|-------------------------|
| Alternia leaf spot      | 40                  | 152                     |
| Bacterial Wilt          | 45                  | 252                     |
| Collar Rot              | 35                  | 202                     |
| Damping Off             | 32                  | 100                     |
Sample of augmented healthy dataset of size 39010 and unhealthy dataset of size 1356 is shown below.

![Healthy augmented dataset](image_url)
7. Discussion

There are two types of image data augmentations, one is called data warping and second is called oversampling. Both of these types are used to improve deep learning model or image classifier. So, these augmentation techniques really help in neural network, machine learning and deep learning applications. Currently there is not any data augmentation technique available that can correct a dataset that has very less diversity. All these augmentation algorithms are performing best under the assumption that the training data and testing data are both taken from the same set of samples.

Following bar graph and line graph shows disease wise number of samples used for augmentation verses number of augmented images for training the classifier.

Following bar graph and line graph shows disease wise number of samples used for augmentation verses number of augmented images for validating the classifier.
8. FUTURE WORK
Once quality dataset with enough volume is ready by performing all above data augmentation techniques, such dataset can be used as part of future work to train the classifier. Such classifiers are used to make classification in many applications in the field of agriculture, healthcare, automotive industries and many more. Even one more step can be done is that, after building and testing classifiers, hyper parameter tuning can be done which improves performance of models. Apart from techniques discussed above, some more techniques like normalization, scaling and preprocessing functions can be used to generate a greater number of images as a part of future work.

9. CONCLUSION
Experimentation done in this research presents series of data augmentation techniques which gives solution to overfitting in machine and deep learning models due to limited size of dataset. In many areas of machine learning, deep learning and big data, this technique gives tremendous benefits to the researchers. This image data augmentation technique is much useful for generating better quality datasets. In coming future, this technique will be used in many applications of emerging areas. Data augmentation is not going to reduce complete noise from input dataset but overall, this is one of recommended technique to create better quality dataset. The major problem of deep learning and machine learning models is overfitting and with the help of these data augmentation technique, the problem of model overfitting can be defeated.

Keywords
Data Augmentation, Machine Learning, Deep Learning, Deep Convolutional Neural Network, Big Data, Image Data Augmentation, Data Science, Image Classification, Machine Learning Classifier, Indian Agriculture, Neural Network

Acknowledgment
We would like to thank the KLE Dr. MSSCET, Belagavi, Karnataka; Visvesvaraya Technological University, Belagavi, Karnataka and Shriram Agriculture College, Mahatma Phule Agriculture University, Solapur, Maharashtra for helping us to carry out this research. Additionally, Opinions, findings, conclusions, or recommendations in this paper are solely of the authors’ and do not reflect the views of the above-mentioned organizations.

References
[1] Connor Shorten and Taghi M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning”, Journal of Big Data, Springer, Shorten and Khoshgoftaar J Big Data (2019)
[2] Atsuhiro Noguchi and Tatsuya Harada, “Image Generation From Small Datasets via Batch Statistics Adaptation”, The University of Tokyo, 2RIKEN, arXiv:1904.01774v4 [cs.CV] 23 Oct 2019
[3] Mr. Balwant J Gorad, Dr. S. Kotrappa, “Comparative Study and Review for Development of Disease Prediction System for Indian Crops”, IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 21, Issue 2, Ser. I (Mar - Apr 2019), PP 68-73
[4] Jason Wang, Luis Perez, “The Effectiveness of Data Augmentation in Image Classification using Deep Learning”, Standford University, Google, 2018
[5] Deep mnist for experts https://www.tensorflow.org/get_started/mnist/prob Accessed: 2017-05-19.
[6] E. Jannik Bjerrum. SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. ArXive-prints, Mar. 2017.
[7] Mr. B J Gorad, Dr. S. Kotrappa, Dr. Jibi Abraham, “Development of Intelligent Monitoring IoT Solution for Indian Agriculture Fields”, NCRET, SITCOE, India
[8] Y. Kubo, G. Tucker, and S. Wiesler. Compacting Neural Network Classifiers via Dropout Training. ArXiv e-prints, Nov. 2016.
[9] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing. Recurrent topic-transition GAN for visual paragraph generation. CoRR, abs/1703.07022, 2017.
[10] Y. Ma and D. Klabjan. Convergence analysis of batch normalization for deep neural nets. CoRR, abs/1705.08011, 2017
[11] M. Marchesi. Megapixel Size Image Creation using Generative Adversarial Networks. ArXiv e-prints, May 2017. 2
[12] B. Wang and D. Klabjan. Regularization for unsupervised deep neural nets. CoRR, abs/1608.04426, 2016
[13] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell. Understanding data augmentation for classification: when to warp? CoRR, abs/1609.08764, 2016
[14] S. Xiang and H. Li. On the Effects of Batch and Weight Normalization in Generative Adversarial Networks. ArXive-prints, Apr. 2017.
[15] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25:1106–14.