IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Augmented Balanced Image Dataset Generator Using AugStatic Library

[1]Allena Venkata Sai Abhishek1, [2]Dr. Venkateswara Rao Gurrala

[1]Research Scholar, [2]Professor

[1]Department of Computer Science and Engineering,

[1]GITAM University, Visakhapatnam, Andhra Pradesh, India

Abstract: The mixed data consists of various structured and unstructured data. The exponential boom of the amount of data has made the datasets of varying samples. This paper focuses on the image dataset generator that balances an imbalanced dataset using the AugStatic augmentation library. The datasets, including various classes, are said to be balanced if the number of samples in the classes is equal. This gives a fair chance for the model to learn about all the classes. An augmented image dataset balances an imbalanced image dataset. It is useful when the data is less in a specific category, generating new data with it. There are multiple augmentation techniques supported by the AugStatic library that helps in developing the augmented balanced library by the iterative implementation of the augmentations on the generated dataset. It takes an input of the existing dataset, majority and minority classes sample count that returns a balanced image dataset by iteratively applying the augmentations on the generated augmented images in the minority class. This generator is efficient and can be used for any image dataset.

IndexTerms **– AugStatic Library, Balanced dataset, Class Imbalance, Dataset Generator, Classification.**

I. INTRODUCTION

The imbalance in the classification dataset is a majorly painful feature of most of the datasets. The class imbalance is removed by over-sampling the minority class. Over-sampling can be done in many ways. Out of which, there are some methods like SMOTE, etc. Exponential increases data in day-to-day have created many complexities for image data. The data augmentation [3] [5] can be done for various data types, such as image, audio, NLP, and Time series.

In this paper, the study is focused on balancing the image dataset using the augmentation library – AugStatic [14]. The balanced dataset helps in giving a fair chance to the model [13] to learn about every class in the dataset. The input for the AugStatic is a NumPy array that returns the augmented images. The count of pictures required for balancing the dataset is given as input, which balances the dataset by generating the input number of samples.

Multiple methods are used for various causes, such as MixUp augmentation [2], which is used for smoothening the images and matting [4], which extracts the image's foreground. Machine learning is a subpart of Artificial intelligence that helps the machine learn from data and help in making a business decision. It learns from a little or no requirement for human mediation. Deep learning is the subpart of machine learning that consists of a complex network that contains layers of nodes. Each node is called a neuron which includes the activation function. Each neuron takes the input with weights and gives an output.

The classification refers to classifying the dataset into various classes. Over the long haul, automation and machine learning have been increasing with the advancement of technology. Data augmentation is a part of data pre-processing [10] that alters the input dataset to amend it into a more suitable format for the train and test sets in a dataset. It helps to mold the data into the model's supported form. Many corporates use machine learning pipelines. And machine learning pipelines include pre-processing steps. Data augmentation generates new synthetic images, which helps mold the data into a required format before feeding it to the model.

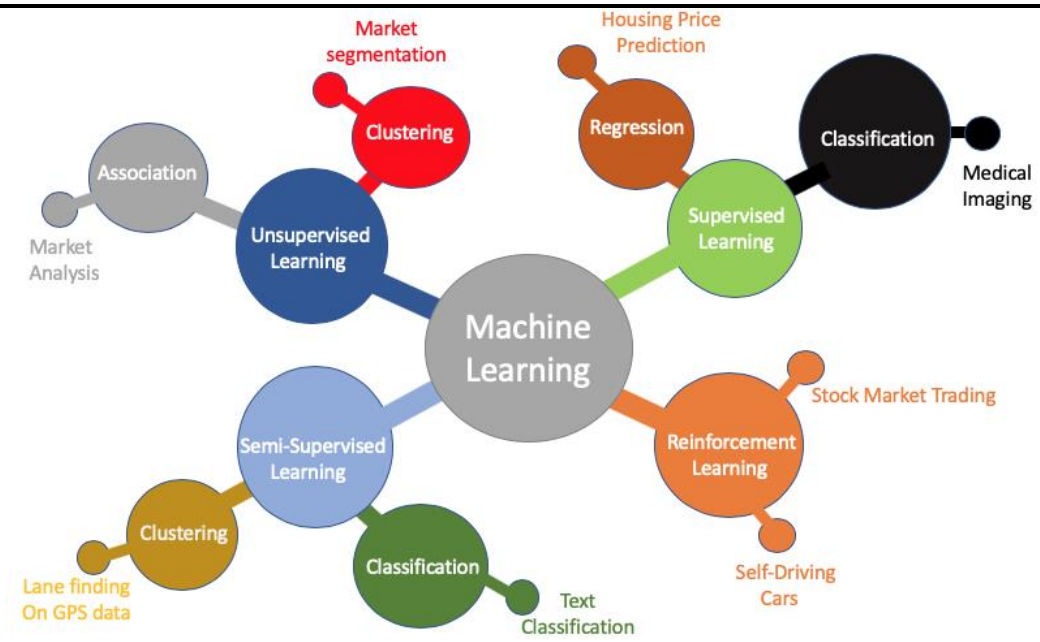

Fig-1: Machine Learning Domains

Data pre-processing is essential for amending the crude data into a cleaned dataset. The data is cleaned by removing the noise, missing values, and different irregularities. Data augmentation is an integral part of the machine learning pipeline. Oversampling increases the number of samples by augmenting the images in the minority class with various augmentation techniques like rotation, scaling, shifting, etc. The image processing pipeline includes pre-processing method. The enhancement and restoration are done in this step, making the data augmentation an integral part of the pipeline.

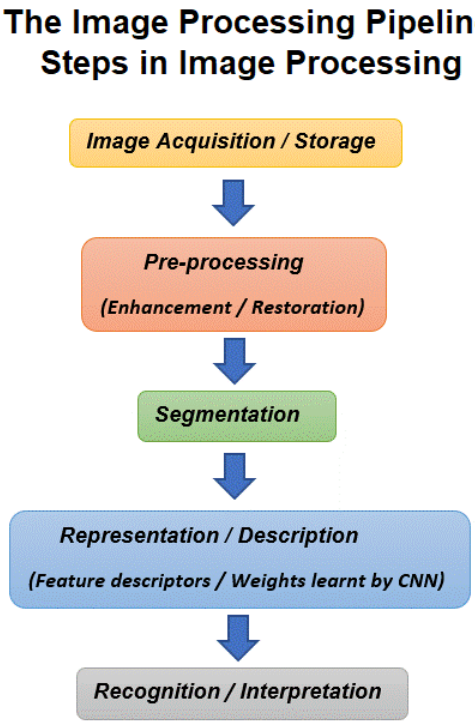

Fig-2: Image processing pipeline

The cost-effectiveness and saving of time make the data augmentation useful. The existing dataset can be oversampled by data augmentation, and the size of the current dataset can be increased; hence, the time for searching for additional data is reduced. Therefore making the augmentation an essential part of many machine learning pipelines.

II. BACKGROUND WORK

Many types of data generators were researched in this study. The workflow of the data generator is shown in fig.3. The data augmentations are applied in a built-in dataset generator that generates the images, which are then fed to the model. The data augmentations can be done for various data types such as audio, text, audio, and time series. There were many augmentation libraries such as Imgaug [11], Albumentations [9], Augmentor [1] and AugStatic. They were extensively compared, and the AugStatic library was chosen to make the augmented balanced dataset generator. Few take input as a NumPy array, and some in tensors [8]. A tensor can be incorporated as a multidimensional array that can be generalized as vectors and matrices. The advantage of using data augmentation is that it is time-saving & cost-effective. Various augmentations were researched and compiled into a compact, lightweight, and practical library. The salient feature of Imgaug, Albumentations, are included in the AugStatic package.

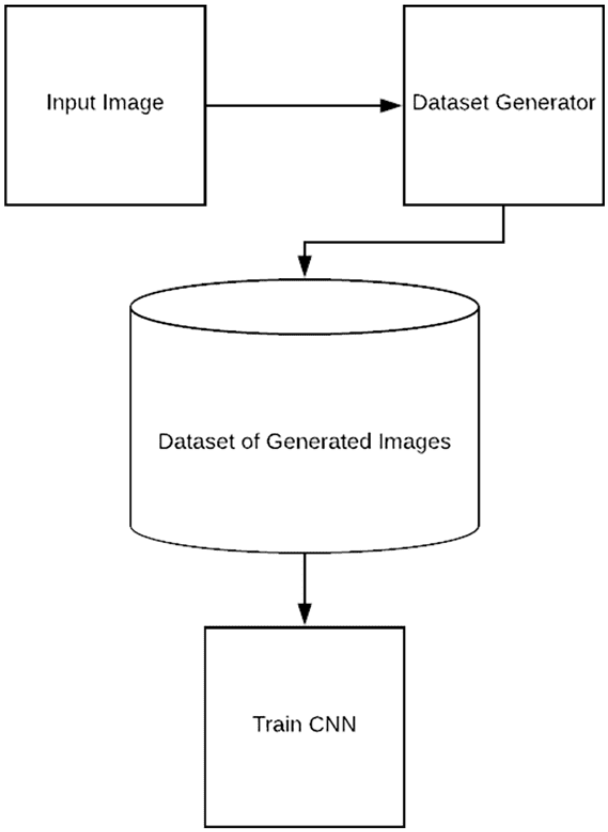

Fig-3: Data Generator

The data generators generate the data. The dataset is balanced by removing the class imbalance. An example of class imbalance is shown in fig. 4.

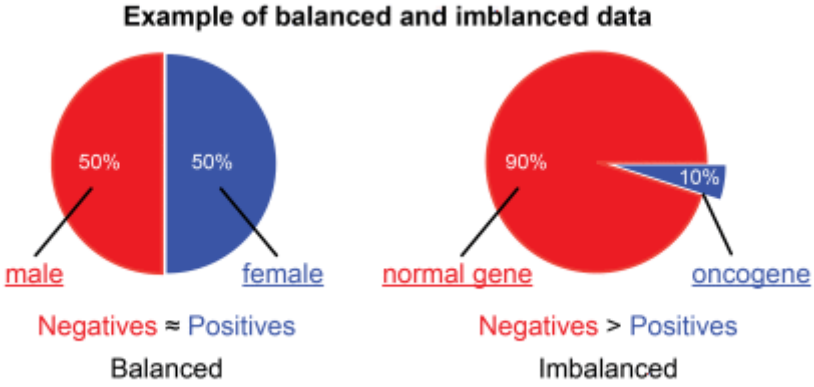

Fig-4: Balanced and imbalanced dataset

There are two types of imbalance – slight and severe inequality. The generator is also used in GANs [6] that generate the synthetic images.

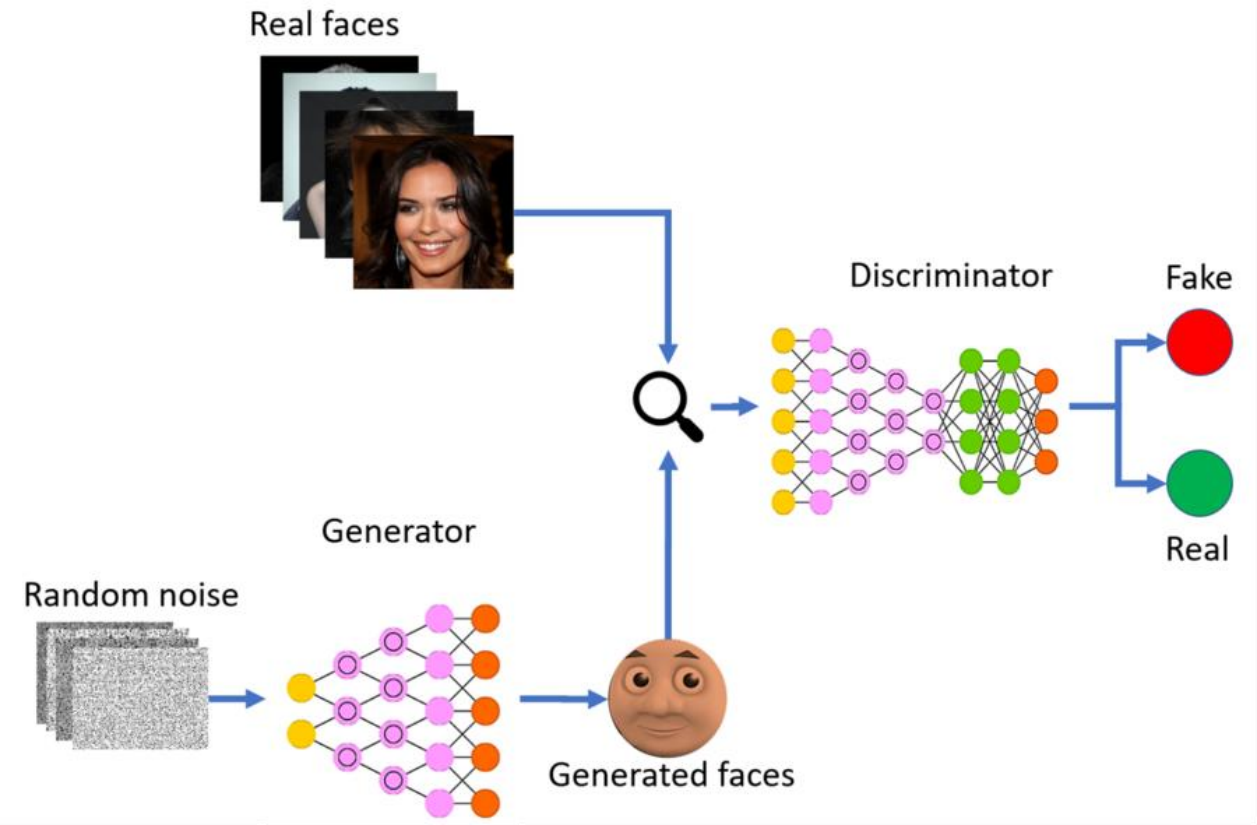

Fig-5: GANs

III. RESEARCH METHODOLOGY

A balanced image data generator is a data generator that is built upon the AugStatic library that takes the sample count in the majority and minority classes as input. The augmented balanced dataset generator is balanced using the augmentation techniques supported by the AugStatic library. The image dataset has various classes. The images of the minority class are fed to the balanced dataset generator. The images are generated by iteratively applying the augmentations on the images and generating new samples and are added up into the minority class. The generated images and the original images are iteratively fed into the generator. The image counts are equalized, making the balanced augmented image dataset. This gives a fair chance for the model to get trained and learn about each class.

The augmentations can be composed and implemented on the images, making the augmentation much more complex and helping the model learn better, as complex augmentation will make it hard for the model to recognize the feature vectors of the images. The object of interest can be better analyzed with a balanced dataset. The generator is built on python, making it easily usable and flexible enough to scale, adding up new features to the generator.

The AugStatic library is advancing with new features making the augmentation techniques various kinds of the augmented balanced dataset for the minority class. The generator is feasible for a multiple class imbalance ratio independent of the type of imbalance. The fed integer value is enough for the generator to work efficiently.

The advantage of a balanced data generator over other data generators is that –

- It removes the class imbalance in a dataset i.e. it balances the number of examples in the classes.
- It helps to increase the data with existing data.
- It can generate various types of augmented dataset as AugStatic supports many types of augmentation transformations.
- Stacked transformations can also be used to generate the balanced augmented dataset.
- Various types of imbalance like slight and severe imbalances can be removed
- The input images are un NumPy arrays, and output images are added up to the minority class in the dataset

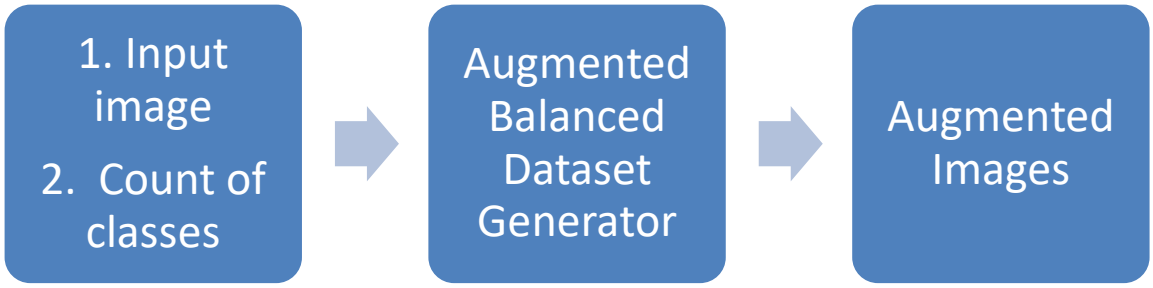

Fig-6: Brief Workflow of Augmented balanced dataset generator

The fast workflow of the augmented balanced dataset generator is shown in Fig. 6. It takes the input of the count of examples in the majority and minority classes in the dataset. The image dataset folders with classes also need to be fed as input to the augmented balanced dataset generator. It asks for the type of augmentation – "single" or "bulk" as input, as shown in Fig. 7.

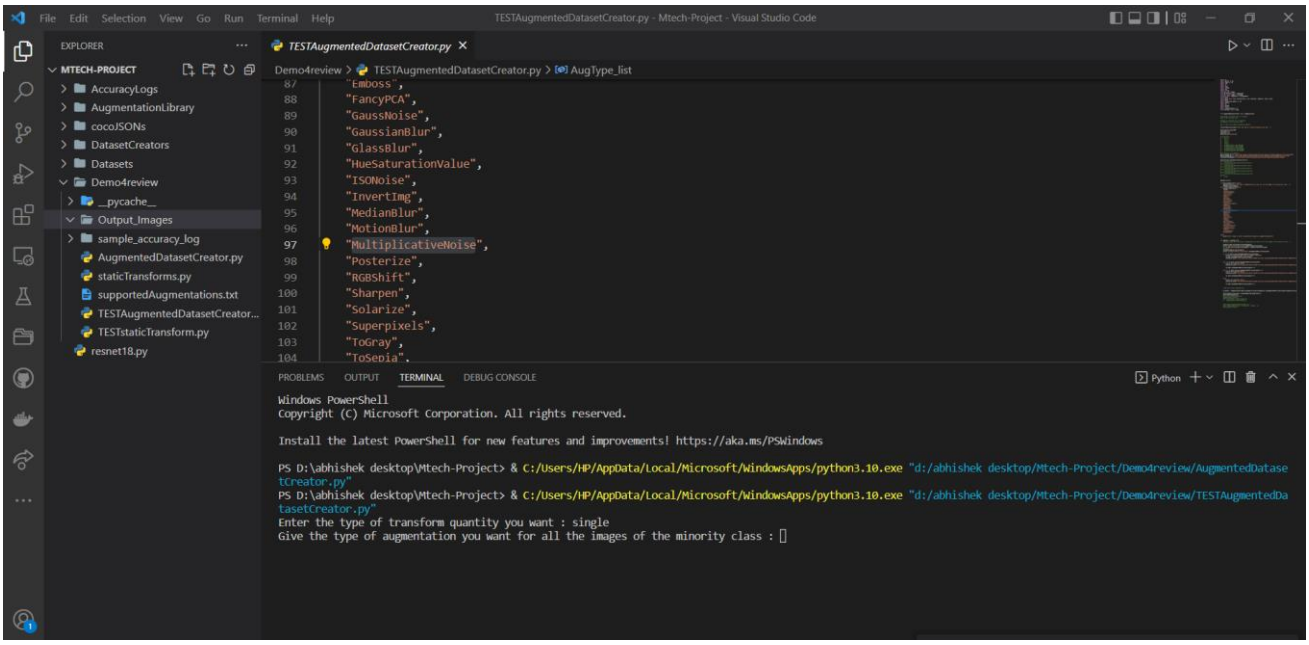

Fig-7: Input -1 of Supported Augmentations by AugStatic library

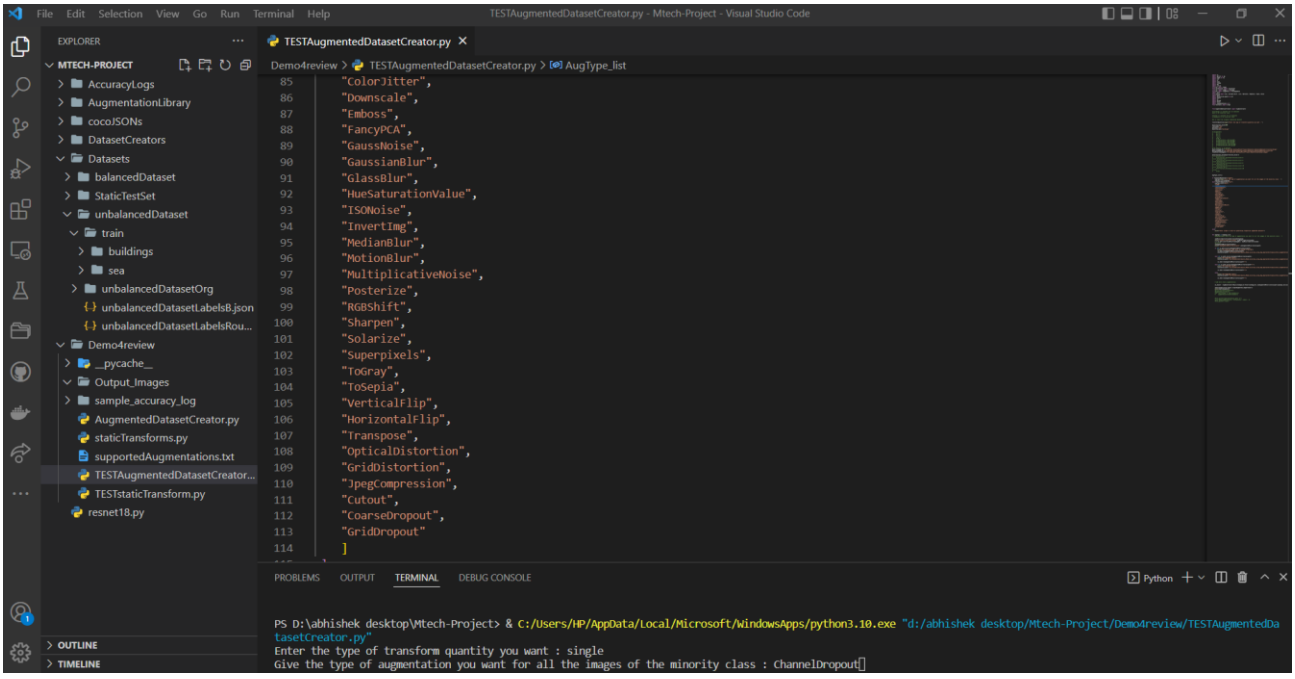

Fig-8: Input -2 of Supported Augmentations by AugStatic library

The second input is taken only for "single." We need to give the type of Augmentation to balance the dataset, as shown in Fig. 8.

For the "bulk" input, there is no need for another input or type of input. It will generate the balanced augmented dataset for all the supported augmentations at a time.

The generator is built on the AugStatic library, supporting all the augmentation techniques in AugStatic. The augmented images are added to the minority class with the existing pictures, henceforth balancing the dataset and removing class imbalance. These types of augmentation techniques supported by the augmented balanced dataset generator are shown in Fig. 9.

Fig-9: Supported Augmentations by AugStatic library

The output of the types of augmentations supported by AugStatic is shown in Fig. 10.

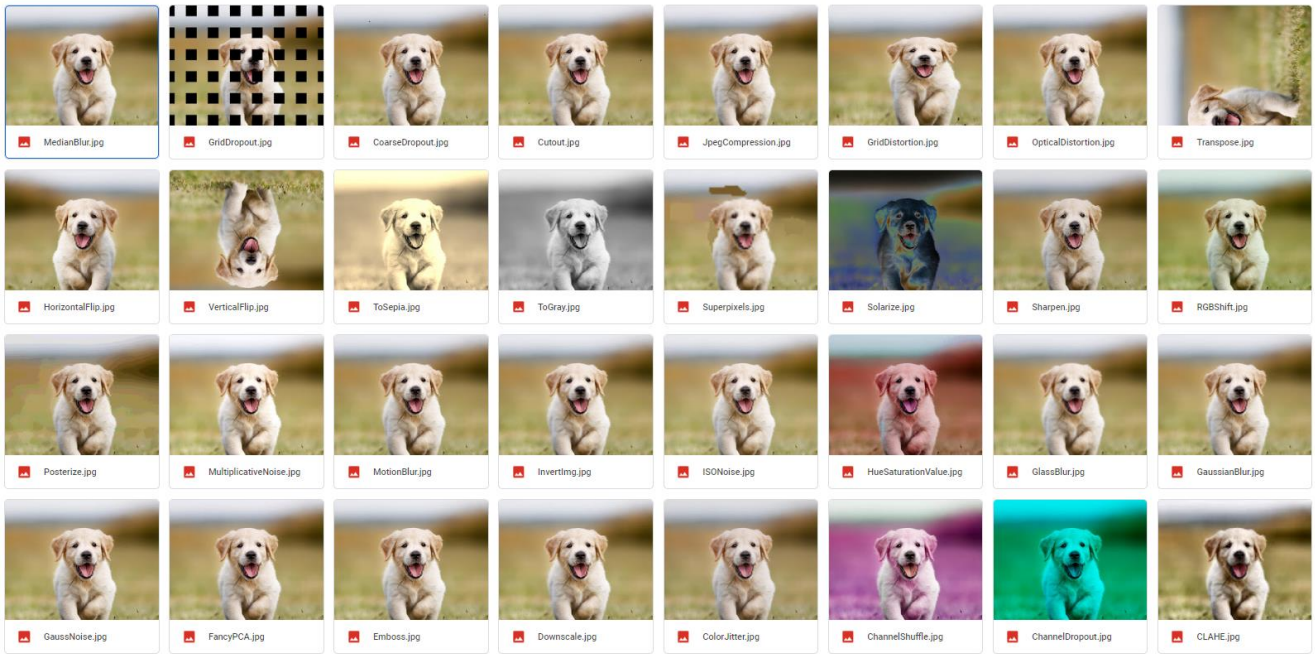

Fig-10: Supported Augmentations by Augmented Balanced Dataset Generator

IV. RESULTS AND DISCUSSIONS

The lightweight and efficient image augmentation balanced dataset generator is developed and explained in this paper. The generator is built upon AugStatic, an efficient augmentation library that supports the PyTorch [7], Keras [12], Imgaug, Albumentations, and Augmentor. The advancement in the features of the library makes the generator scalable.

It takes the classes' image count and the dataset folder with the class name. as input. Then the augmentation type is applied as input generating the balanced augmented dataset as shown in Fig. 11.

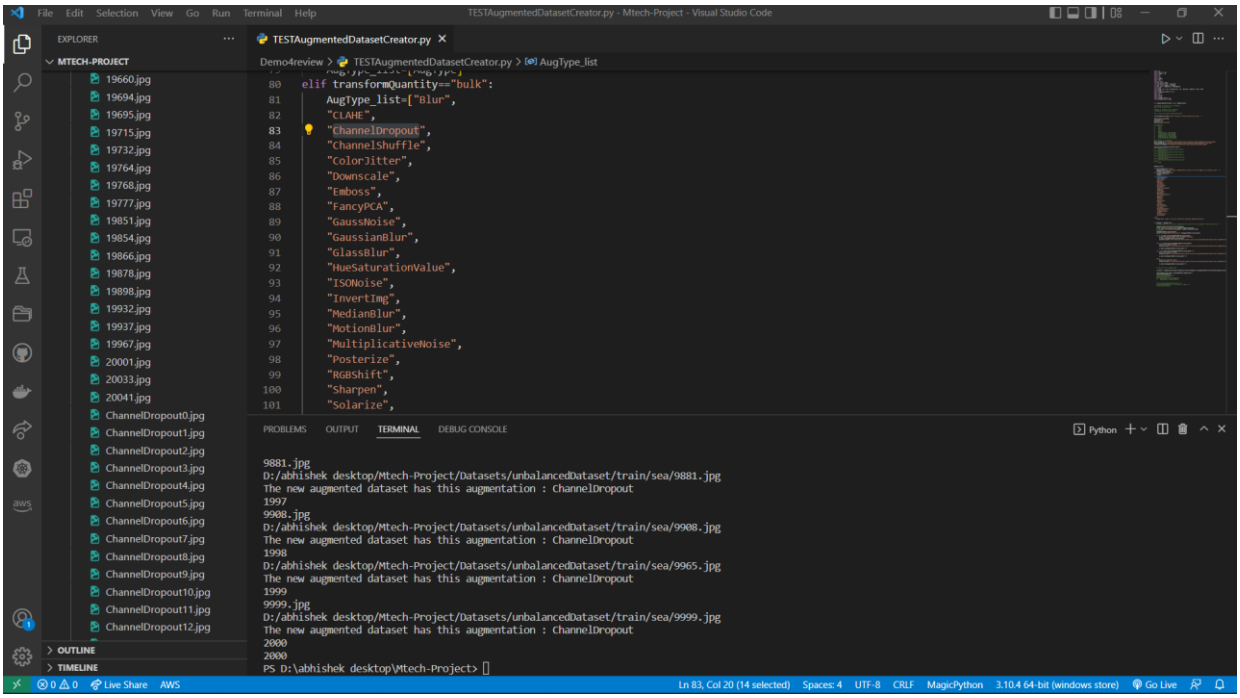

Fig-11: Output of Supported Augmentations by AugStatic library

V. CONCLUSION AND FUTURE SCOPE

Removal of class imbalance by balancing the dataset using data augmentation is an essential feature for a machine learning pipeline to give the machine learning model a fair chance to learn about all the classes in an image dataset. It is built upon the AugStatic augmentation library, making it efficient and scalable with features in the AugStatic library.

The scaling of the generator is dependent upon the advancement of the features of the library, and the random augmentations will be added to the AugStatic library. The class imbalance can help save time in searching for data. The AugStatic is scalable, making the generator scalable. The scalability is explained and demonstrated in this paper. With the advancement in augmentation, there is a lot of scope in making the generator for audio, NLP, and time-series data.

REFERENCES

[1] Marcus D. Bloice, Christof Stocker, Andreas Holzinger. 2017. "Augmentor: An Image Augmentation Library for Machine Learning" arXiv:1708.04680v1 [cs.CV]

[2] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, David Lopez-Paz. 2018. "MixUp augmentation for image classification"arXiv:1710.09412v2 [cs.LG]

[3] Zhiting Hu, Bowen Tan, Ruslan Salakhutdinov, Tom Mitchell, Eric P. Xing. 2019. "Learning Data Manipulation for Augmentation and Weighting " arXiv:1910.12795v1 [cs.LG]

[4] Shanchuan Lin, Linjie Yang, Imran Saleemi, Soumyadip Sengupta. 2021. "Robust High-Resolution Video Matting with Temporal Guidance" arXiv:2108.11515v1 [cs.CV]

[5] Connor Taghi M. Khoshgoftaar. 2019. "A survey on Image Data Augmentation for Deep Learning" Shorten and Khoshgoftaar J Big Data

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio "Generative Adversarial Networks " arXiv:1406.2661v1 [stat.ML]

[7] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, Soumith Chintala. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library" arXiv:1912.01703 [cs.LG]

[8] Dimitrios Koutsoukos, Supun Nakandala, Konstantinos Karanasos, Karla Saur, Gustavo Alonso , Matteo Interlandi "Tensors: An abstraction for general data processing."

[9] Alexandr A. Kalinin, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Alexander Buslaev. 2020. "Albumentations: fast and flexible image Augmentations"

[10] Sujith Jayaprakash Balamurugan E. 2015. A Comprehensive Survey on Data Preprocessing Methods in Web Usage Mining.

[11] "Augmentation methods applied to data using imgaug library.", https://github.com/aleju/imgaug

[12] Ketkar, Nikhil. (2017). "Introduction to Keras." 10.1007/978-1-4842-2766-4_7.

[13] Allena Venkata Sai Abhishek, Sonali Kotni, 2021, Detectron2 Object Detection & Manipulating Images using Cartoonization, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 08 (August 2021),

[14] Allena Venkata Sai Abhishek, Dr. Venkateswara Rao Gurrala. May-2022. "AUGSTATIC - A LIGHT-WEIGHT IMAGE AUGMENTATION LIBRARY ", International Journal of Emerging Technologies and Innovative Research (www.jetir.org | UGC and issn Approved), ISSN:2349-5162, Vol.9, Issue 5