IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.6, Issue.10, Oct. 2018 E-ISSN: 2347-2693

Efficient Learning on Imbalanced Image Set

Shivani Guldas

Dept. of ISE, Ramaiah Institute of Technology, Bangalore-54, Karnataka, India

**Corresponding Author: shivanikguldas@gmail.com, Tel.: +91-9036030780*

Available online at: www.ijcseonline.org

Accepted: 7/Oct/2018, Published: 31/Oct/2018

Abstract— Handling imbalanced image sets is a challenging issue being faced by the conventional categorizer. Imbalance problem occur with real world data due to various reasons, to which the ordinary classifiers gets influenced towards major class data. In this paper, we aim to balance bi-class absolute image set by creating synthetic samples of minority class images. Tests on three image sets using five synthetic image generation methods, four image features and three evaluation measures is carried out. KNN classification is performed on all three image set which are pretty imbalanced and the results indicate that synthetic creation of minor class images progresses the performance measures.

Keywords— Imbalanced image set, k-nn categorization, Synthetic image generation, performance measures improvement

I. Introduction

Imbalanced data is one of the challenging problem the world is facing today in real life and a research in this field has gained a rapid pace. Balancing the dataset helps in efficient analysis in various fields like medical, agriculture, industrial system monitoring, behavior analysis, activity recognition etc[1].The data is said to be imbalanced when one class is well represented than other class, the imbalance problem is divided into two types: absolute ratio and the relative ratio. When the ratio between majority and minority class is not equal is called relative imbalance type and when there are less minority samples is called absolute type. This will lead to inefficient analysis in many fields specified above and also degrade the performance of conventional categorizers. The most compulsive one is a minor class and in general, the categorizer will be more influenced by major class. What factors does cause imbalance problem is one fine question that strike when it comes to solve the imbalance issue; the imbalance ratio (minority instances v/s majority instance), overall training size, complexity of data, categorization technique and assumption of many categorizers that data is balanced are the factors that lead to imbalance issue[9]. To tackle with imbalance data, we have three approaches [3]:

Data level approach:
This approach focus on altering the image set and help conventional categorizer in learning efficiently. Oversampling and under sampling are common approaches that help in balancing the distribution.

Algorithm level approach:
This approach deal with modifying existing learning algorithms to ease the favouritism towards majority class and acclimate them to mine data with skewed distributions.

Hybrid approach:
This approach is a combination of both data and algorithm level approach where the algorithm is altered clubbed with other or arrive at a new algorithm. This approach also focus on under sampling / oversampling methods with the modified algorithm.

This paper works on analysing how well the conventional categorizer perform while dealing with data-level approach for image set? We consider KNN classification for balancing bi-class image set with proper collection of fine image samples of both classes with similar properties. Feature selection of image for categorization play important role in image categorization [8]. The project prioritizes data-level method on bi-class image sets. Our approach is to balance imbalanced image set by creating synthetic samples of minority class images [6]. By using synthetic image generation techniques with respect to absolute imbalance type, we can create synthetic samples and train the categorizer in order to progress performance measures. As per the survey, oversampling is better than the under sampling technique when dealing with data-level method [9][1].

The paper is systematized as follows, section 2 describes literature survey, section 3 explains Synthetic picture generation techniques used, features considered and the performance measures estimated. Section 4 consist of experimentation and its results with different image set, section 5 deal with conclusion and future work.

II. LITERATURE SURVEY

Bartosz Krawezyk [3] , discuss about the current challenges the researchers are facing today with respect to imbalanced data in all real time applications and also guides with future directions to deal with data imbalance. He also discussed methods to tackle with imbalanced data, open challenges in binary and multi-class classification, regression etc. Apurva Sonak and R A Patankar[9] describes what imbalance problem is, the factors affecting the datasets and methods to deal with the imbalanced data. They categorize the solution into two categories: Sampling and cost-sensitive learning. Under sampling we have over-sampling and under sampling. The comparison between methods is made and conveys that under sampling outperforms oversampling while dealing with data-level approach.

Dr. D Ramyachitra and P Manikandan[12] discusses the characteristics of imbalanced problem, the methods to deal with it. They describe the algorithms, the type of approaches that one can deal to solve imbalance issue and the fields that face imbalance issue. They also describe the evaluation matrix of imbalance issue and finally convey that data-level approach-oversampling technique is best to solve the imbalance issue.

Feature extraction play an important role when dealing with pictures as collection of pixels form a picture. Features provide the information to solve the issue and refer to structures from simple to complex types [8]. Ruchika Mishra and Utkarsh sharma [15] discuss the two approaches for synthetic sample creation: Spatial domain and frequency domain and they compare between two. Advantages and disadvantages are also discussed for both the approaches and they find it hard to convey which is the best technique. Anu Namdeo and Sandeep Singh Bhadoriya [6] describe different synthetic image generation techniques for the better quality images and describe each method and their advantages and disadvantages.

We need to consider standard measures in order to evaluate the classifier performance, D. Druga Prasad and Dr. K Nageswar Rao [1] arrived at a novel algorithm WIMOTE which handles imbalanced dataset and improves performance measures. The technique follows oversampling for minor set by synthetic generation of data and under sampling in major set data. They experimented on 15 different datasets and convey the results. The WIMOTE method improve the performance measures well- accuracy, precision, recall, F-measure, AUC). We also measure specificity in our project. With the survey done, our approach in improving performance is explained in further sections.

III. SYNTHETIC PICTURE GENERATION TECHNIQUES

When the data is limited, the minority set pictures have a very high error categorization rate for an imbalanced picture set. To overcome this, synthetic images are generated in order to balance the classes. Generation of synthetic images and building an unbiased classifier model includes the following steps.

Step 1: Find the first order statistical features of the minority training samples
Step 2: Generate synthetic instances using these features by using Synthetic picture generation techniques
Step 2: Append the synthetic instances to the modified Training set
Step 3: Run the nearest neighbor classifier over the updated training data.

Statistical Feature Extraction:

In picture classification, selection of features play an important role. The features are categorized into two types-texture or statistical, structural [19][20]. The property of picture that represent the surface and structure of a picture or pattern of a picture [20]. The features extracted for this project are explained below.

Mean:

The mean is defined as an average of all pixel values of a picture. It can be calculated by the formula given below

$$\mu = \frac{1}{MN}\sum_{i=0}^{M}\sum_{j=0}^{n} p(i,j) \qquad \text{eq(1)}$$

where 'p' represent pixel value at point i, j for MxN sized picture [19][8].

Standard deviation:

This normalizes the facts and tell how far our data is from the mean. Larger the standard deviation value farther is the mean. If smaller standard deviation, the data is closer to mean. It is a portion of the spread of a set of values from the average value. It also represents dispersion of local regions. This can be calculated by the formula given below

$$\sigma = \sqrt{\frac{1}{MN}\sum_{i=0}^{M}\sum_{i=0}^{n}(p(i,j)-\mu)^2} \qquad \text{eq(2)}$$

which represent root of pixel 'p' at points i, j for MxN picture from its mean [19][8].

RMS (Root Mean Square):

Alias quadratic mean is a statistical measure of the degree of a fluctuating quantity. It is useful when variates are positive and negative. It is the root of mean square value of each row or column or of an entire picture. It can be calculated by the formula given below

$$Y=\sqrt{\sum_{i=0}^{n}\left[u_{ij}\right]^{2}}\Big/_{M} \qquad \text{eq(3)}$$

Entropy:

It is a measure of randomness of a picture used to categorize the surface of a picture. Higher the entropy, higher, the higher mis-categorization. It is a distribution variation in an area. It can be represented as [8][19]

$$h = -\sum_{k=0}^{L-1} P_{rk}\,(log_2 P_{rk}) \qquad \text{eq(4)}$$

All the above features discussed are embattled to discriminate between pictures of different classes.

Techniques:

Synthetic picture generation consist of two types: Spatial domain and Frequency domain, where Spatial domain deal with direct manipulation of pixels of an image and Frequency domain works on Fourier transform of image in order to transform an image [15]. In this work we work with spatial domain techniques.

Contrast stretch:

The method tries to progress a picture by elongating the series of strength values it contains in order to utilize possible values. Contrast extending is constrained to a lined plotting of input to output values [6].

Histogram Equalization:

This technique is used to boost the look of dark pictures. The dark picture histogram would be crooked to the minor tip of the grey measure, in order to compact all the picture detail towards the dark tip of the histogram. If the grey levels are 'expensed out' at dark tip to produce a more uniform dispersed histogram then the picture will be more unblemished. This method elasticities the histogram across all the pixels (0 – 255). It is mainly used to increase the contrast of pictures for the definiteness of human review, normalizes lighting changes besides help in image understanding problems [6].

Contrast enhancement:

This technique inevitably revitalizes unclear or dark images, apply proper tone correction to gain better picture. It helps mainly in medical field- to capture internal structure of human, to check fractures of bone (X-ray). For example, X-ray generates nil contrast picture because of water presence in human body, at such time this technique help in acquiring a clear picture [6].

Brightness preserving Bi-Histogram equalization:

As brightness preserving is an important feature of an image, this method is used to shield the illumination of an image by splitting the picture's histogram into two equal parts in order to make sure that the strengths are also equally arranged [6].

Adaptive histogram equalization:

Versatile HE is mainly useful in improving divergence as a piece of pictures. It varies from Histogram Equalization by malleable method that figures a couple of histograms and each histogram recognizing with a specific portion of a picture. The difference region for a picture won't be satisfactorily improved by Histogram Equalization. [6]The method helps this advancement by adjusting each pixel with a alteration limit got from a region district. It is used to vanquish a couple of checks of overall direct min-max windowing system. Thusly it diminishes the proportion of bustle in areas of the picture. Besides, this technique have the limit with respect to upgrading the distinction of grayscale and shading picture.

Performance Measures:

In this experiment, we estimate specificity, recall and precision and consider the following elements to compute them. In this project, the terms TP (true positive), TN (true negative), FP (false positive) and FN (false negative) is defined as follows:

- *TP= Positive pixels correctly recognized as positive*
- *TN= Negative pixels correctly recognized as negative*
- *FP= Negative pixels recognized as positive*
- *FN= Positive pixels recognized as negative*

Specificity:

It is the ratio of negative pixels identified as negative out of total number of negatives. It is also known as True negative rate and is assessed as:

Specificity = $\frac{TN}{TN+FP}$……………………..…………eq (5)

Recall:

It is the ratio of correctly predicted positive pixels as positive pixels out of all positive pixels. It is also known as TPR (true positive rate or sensitivity) and is assessed as:

Recall = $\frac{TP}{TP+FN}$…………………….. ……………….eq (6)

Precision:

The ratio of correctly identified as positive pixels as positive out of all positive predictions. It is also known as PVV (positive prediction value) and is assessed as:

Precision = $\frac{TP}{TP+FP}$……………………………………eq(7)

IV. Experiment And Results

Proposed Approach:

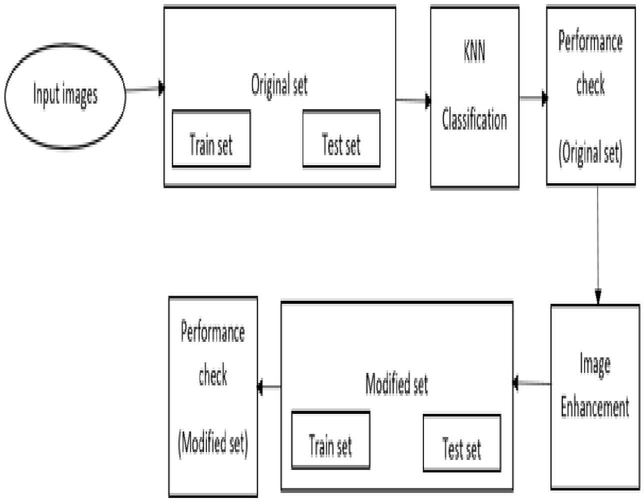

Figure 1: Proposed approach for imbalanced image set

In this approach, we provide different bi-class image sets as inputs and consider some images for training and testing. First stage of project deals with applying KNN classification, categorize the test data for binary class and check the performance of classifier considering test images of original set. Then consider same set of images stored in modified set on which synthetic picture generation techniques are applied to minor class images and store them in same (modified train set). Now, calculate the performance measures after modified set is trained by nearest neighbour and check whether there is an improvement in measures or not. We implement this using MATLAB scripting- MATLAB 2013/MATLAB 2017 tool.

The approach followed is a data-level and we aim to improve performance measures by creating synthetic images of minor class with five different techniques. The results are discussed below with three different image sets.

Results:

Image set description: Image set 1- The first image set considered is the forest and highway where the upper class is the highway class and the lower class is forest. The images are considered from the www.pexels.com website with similar properties.

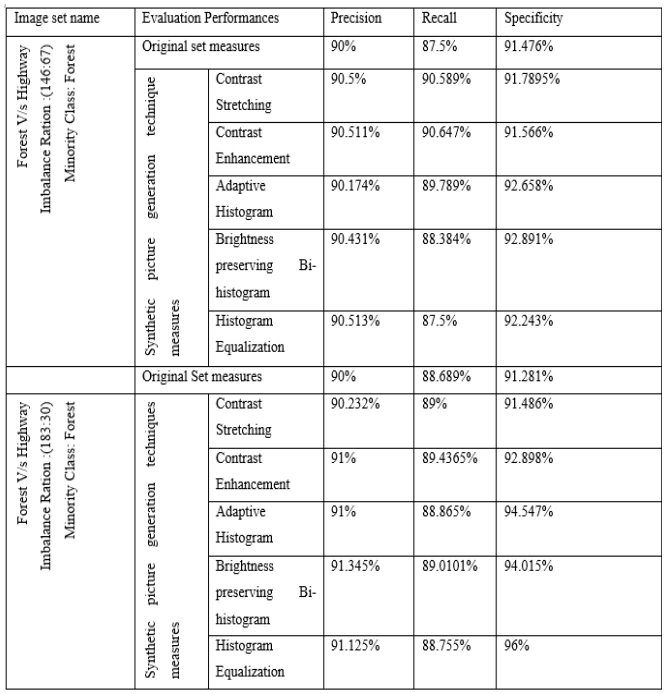

Image set name	Evaluation Performances		Precision	Recall	Specificity
Forest V/s Highway Imbalance Ration :(146:67) Minority Class: Forest	Original set measures		90%	87.5%	91.476%
	Synthetic picture generation technique measures	Contrast Stretching	90.5%	90.589%	91.7895%
		Contrast Enhancement	90.511%	90.647%	91.566%
		Adaptive Histogram	90.174%	89.789%	92.658%
		Brightness preserving Bi-histogram	90.431%	88.384%	92.891%
		Histogram Equalization	90.513%	87.5%	92.243%
	Original Set measures		90%	88.689%	91.281%
Forest V/s Highway Imbalance Ration :(183:30) Minority Class: Forest	Synthetic picture generation techniques measures	Contrast Stretching	90.232%	89%	91.486%
		Contrast Enhancement	91%	89.4365%	92.898%
		Adaptive Histogram	91%	88.865%	94.547%
		Brightness preserving Bi-histogram	91.345%	89.0101%	94.015%
		Histogram Equalization	91.125%	88.755%	96%

Table 1: Synthetic sample generation for forest v/s highway image set

Image set description: Image set 2- The second image set considered is cats v's dogs where the upper class is cat's class and the lower class is dog's. The images are considered from the www.imagenet.com website with similar properties.

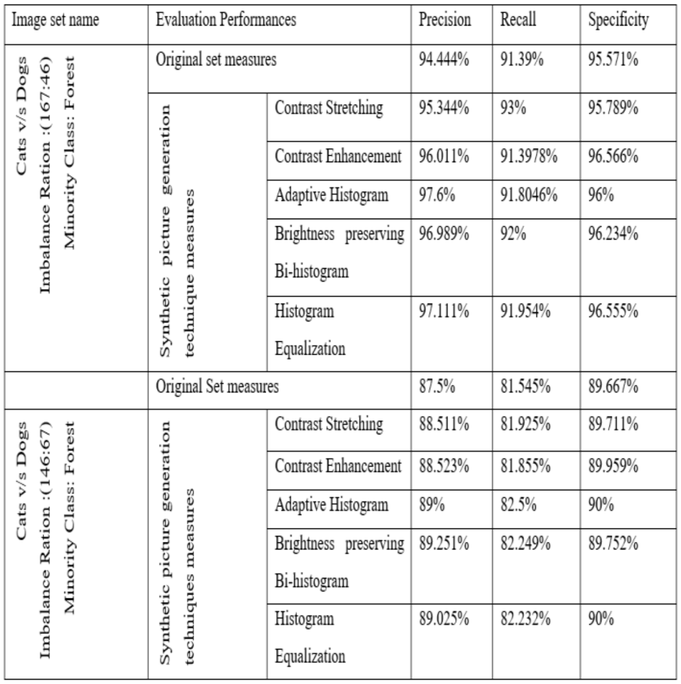

Image set name	Evaluation Performances		Precision	Recall	Specificity
Cats v/s Dogs Imbalance Ration :(167:46) Minority Class: Forest	Original set measures		94.444%	91.39%	95.571%
	Synthetic picture generation technique measures	Contrast Stretching	95.344%	93%	95.789%
		Contrast Enhancement	96.011%	91.3978%	96.566%
		Adaptive Histogram	97.6%	91.8046%	96%
		Brightness preserving Bi-histogram	96.989%	92%	96.234%
		Histogram Equalization	97.111%	91.954%	96.555%
	Original Set measures		87.5%	81.545%	89.667%
Cats v/s Dogs Imbalance Ration :(146:67) Minority Class: Forest	Synthetic picture generation techniques measures	Contrast Stretching	88.511%	81.925%	89.711%
		Contrast Enhancement	88.523%	81.855%	89.959%
		Adaptive Histogram	89%	82.5%	90%
		Brightness preserving Bi-histogram	89.251%	82.249%	89.752%
		Histogram Equalization	89.025%	82.232%	90%

Table 2: Synthetic sample generation for cat's v/s dog's image set

Image set description: Image set 3- The third image set considered is garbage v/s clean streets where the upper class is clean street class and the lower class is garbage. The images are considered from the www.pexels.com website with similar properties.

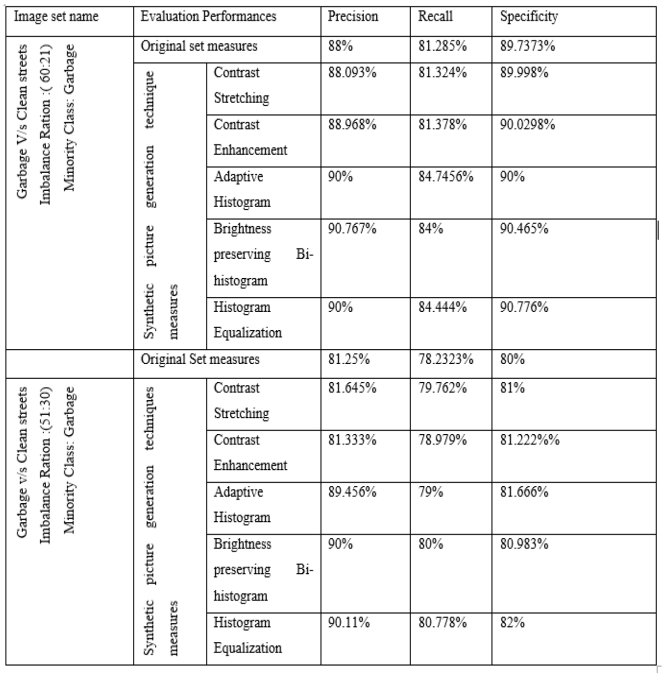

Image set name	Evaluation Performances		Precision	Recall	Specificity
Garbage V/s Clean streets Imbalance Ration :(60:21) Minority Class: Garbage	Original set measures		88%	81.285%	89.7373%
	Synthetic picture generation technique measures	Contrast Stretching	88.093%	81.324%	89.998%
		Contrast Enhancement	88.968%	81.378%	90.0298%
		Adaptive Histogram	90%	84.7456%	90%
		Brightness preserving Bi-histogram	90.767%	84%	90.465%
		Histogram Equalization	90%	84.444%	90.776%
	Original Set measures		81.25%	78.2323%	80%
Garbage v/s Clean streets Imbalance Ration :(51:30) Minority Class: Garbage	Synthetic picture generation techniques measures	Contrast Stretching	81.645%	79.762%	81%
		Contrast Enhancement	81.333%	78.979%	81.222%%
		Adaptive Histogram	89.456%	79%	81.666%
		Brightness preserving Bi-histogram	90%	80%	80.983%
		Histogram Equalization	90.11%	80.778%	82%

Table 3: Synthetic sample generation for garbage v/s clean streets image set

Observations:

We observe that the above five techniques discussed help in improving the measures after modifying the minor class pictures of training set and there is positive outcome for few methods for few of the measures. In the first image set, forest v/s highway (146:67) - there is improvement in measures with all the techniques apart from histogram equalization. For imbalanced ratio (183:30), there is improvement in all the techniques. The second image set cats v/s dogs (167:46) ratio there is improvement in all techniques and the ratio-(146:67) there is improvement in recall and specificity except precision in two of the methods i.e, adaptive histogram and Brightness preserving Bi-histogram. The third image set, Garbage v/s clean roads (60:21) and (51:30), there is improvement with all techniques discussed. By this approach we infer that for few techniques if recall is not improved but recall and specificity is improved.

V. CONCLUSION AND FUTURE WORK

In this paper we intended to improve the system performance measures like precision, recall and specificity of imbalanced image set. The proposed approach uses KNN classifier to categorize the image set, create synthetic samples of minor class images using synthetic image generation techniques and retrain the system with modified set of images. The classification is performed on three different binary class image sets and the experiment shows improvement. The result convey that at data-level approach, synthetic generation of minor class pictures improves the measures-precision, recall and specificity. In future, we would like to deal with more intricate image sets and many other synthetic image generation techniques.

REFERENCES

[1] Dr. Durga Prasad, Dr. K Nageswar Rao, “An Imprved approach on class imbalance data using within class minority oversampling technique” International Journal of latest trends in engineering and technology, VOL.7, 2017l.

[2] Yingying Qin, Wenjie Chen, Jie Chen, “Generating images for imbalanced dataset problem” Institute of Electrical and Electronics Engineer (IEEE), 2016.

[3] Bartosz Krawezyk, “Learning from Imbalanced data: open challenges and future data”, 2016, Springerlink.com.

[4] Jiaojiao Li, Qian Du, Wei Li, Yunsng Li “Representation based yper spectral Image Classification with Imbalanced data”, Institute of Electrical and Electronics Engineers (IEEE), 2016.

[5] Gregory Luppescu, Raj Shah,“Personalized Image Enhancement” Institute of Electrical and Electronics Engineers(IEEE) , 2016.

[6] Anu Namdeo, Sandeep Singh Bhadriya, “A Review on Image enhancement techniques with its advantages and disadvantages” IJSART, 2016.

[7] Hebatallah Mostafa Anwer, Mohamed Farouk, Ayman Abdel-HamAida Ali, Siti Mariyam Shamssuddin, Anca L Ralescu, “Classification with class imbalance problem: A review” Intertiol Jourl Of Advances In Soft Computing And Its Applications, 7 (3). pp. 176-204, 2015.

[8] Vaishnavi L Kaundanya, Anita Patil, Ashish Panat “Classfication of Emotions from EGG using KNN classifier” 2015, IJSET.

[9] Apurva Sonak, R A Patankar “A survey on methods to handle Imbalance dataset” 2015 International Conference on Computing, Networking an d Communications(IJCSMC).

[10] Tdd Perry, Mhommed Bader-El-Den, Steven Cooper, “Imbalanced classification using Genetically Optimized Cost sensitive classifiers” Institute of Electrical and Electronics Engineers (IEEE),2015.

[11] M Srinivas, R Bharath, P Rajalakshmi, C Krishna Mohan, “Multi-level categorization method for medical methods”, 2015, 17th International conference n E-Health Networking, Application & Services (HealthCom).

[12] Dr. D Ramyachitra, P Manikandan,“ Imbalanced dataset classification and solutions: A Review” July 4 2014, International Journal of Computing and Business Research (IJCBR).

[13] M Akkil Jabbar, B L Deekshatulu, Priti Chandra “Classification of Heart disease using KNN and Genetic algorithm” International Conference on Cmputational Intelligence: Modelling Techniques and Applications(CIMTA,)2013.

[14] Mr. Rushi Lngadge, Ms. Snehalata S Dngre, Dr. Latesh Malik, “Class Imbalance problem in data mining: A Review”, International Journal of Computer Science and Network (IJCSN), Feb 2013 .

[15] Ruchika Mishra, Utkarsh Sharma, “Review of Image Enhancement Technique”, 7th International Journal of Engineering Research and Technology (IJERT), August 8 2013.

[16] Nour Moustafa, Jill Slay "Improving classification performance for the Minority class in highly imbalanced dataset using Boosting", 2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12).

[17] Arun Kumar M N, H S Sheshadri, "On the class of Imbalanced datasets", International Journal of Computer Applications (IJCA), April, 2012.

[18] Tarek M Bittibssi, Gouda I Salama, Yehia Z Mehaseb and Adel E Henawy, "Image Enhancement Algorithms using FPGA", 2012 8th International Computer Engineering Conference (ICENCO).

[19] Pradeep N, Girisha H, Sreepathi B and Karibasappa K, "Feature extraction of Mammograms", International journal of -l of Bioinformatics research, Vol 4, 2012.

[20] G N Srinivasa, Shoba G, "Statistical Texture analysis", Proceedings of world academy of Science and Technology-vol 36, dec 2008.

[21] Wacharasak Siriseriwan and Krung Sinapiromsaran, " The Effective Redistribution of Imbalance Dataset: Relocating safe-level SMOTEwith Minority outcast handling", Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok 10300, Thailand, 12 Nov 2014.

Authors Profile

Shivani Guldas, She received B.E degree in Information Science and Engineering from B V Bhoomaraddi college of Engineering &Technologies. She is currently pursuing M.Tech in Software engineering at Ramaiah Institute of Technology, Bengaluru. Her research interest is in Software engineering, Software testing Artificial Intelligence and Machine Learning.