**CHAPTER 1**

# 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, behaviour analysis, activity recognition etc. The imbalance problem is divided into two types: absolute ratio and the relative ratio. Factors that cause imbalance problem is one fine challenging thought that strikes when dealing with imbalance issue. The factors that lead to imbalanced issue are 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.

To tackle with imbalance data, there are three basic approaches:

***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 [3].

***Algorithm level approach:***

This approach will modify existing learning algorithms to ease the favouritism towards majority class and acclimate them to mine data with skewed distributions [3].

***Hybrid approach:***

This approach is a combination of both data and algorithm level approach where the algorithm is altered, or arrive at a new algorithm where two different algorithms are clubbed to come up with solution. This approach also focus on under sampling / oversampling methods with the modified algorithm [3].

In this project, data- level approach has been considered to handle imbalanced image set. As per the survey, oversampling is better than under sampling technique when dealing with data-level method [9][1]. Hence Oversampling approach is considered in this project to balance the bi class image set.

Considering bi-class image set with fine collection of images of two classes with similar effigy properties, the statistical features are calculated for every depiction. KNN categorization is performed based on the features calculated as the feature selection of images play important role in image categorization [8]. The performance is evaluated using different evaluation measures. Synthetic copies of minor class images are created using synthetic image generation techniques. The system is retrained in order to improve the performance of the system and performance is rechecked [6].

**1.1 Problem Statement**

Many researchers, from decades are trying to resolve imbalanced issue, yet have failed to achieve noteworthy performance. Categorizing the normal dataset (text) is easier to that of categorizing image set. There involves lot of challenges, as discussed earlier above. The first challenge was in collecting an imbalanced image set with similar properties of image, choosing the categorizer and categorising the image set. The next challenge involved generating synthetic pictures in order to balance the image set by enhancing the image of interested (minor) class using image enhancement techniques. Third challenge is to improve the performance measures.

### 1.2 Motivation

When one wants to carry out analysis on image sets and if the image set is imbalanced or if there are no sufficient images for a category to perform analysis, then one cannot analyse well and it would be tedious to come up with any outcome. As this is the current issue (imbalanced class), it has to be resolved in order to make proper analysis and come with a best solution to overcome the imbalanced issue. To come up with a good outcome, having a balanced image set is essential. From decades, researchers are trying to balance the imbalanced data by various approaches.

### 1.3 Scope and Objective

Many fields like agriculture, healthcare, medical, etc make use of image as their data, but the data would be irrelevant or missing or unlabelled; in medical area, rare diseases may have very less images which leads to improper dataset. When the instances of one class are less than other class then the data is imbalanced and it has to be balanced. In order to balance the data, there are different approaches that can be followed in order to improve system performance.

The objectives of the project is as follows:

* + Collect imbalanced image set
  + Categorizing the images using classification techniques
  + Feature selection
  + Synthetic image generation
  + Improve the performance of classifier

### Organization of the Report

The report is organized as follows:

The chapter 1 comprise introduction followed by literature survey as Chapter 2. Chapter 3 consist of Requirement analysis, functional and non-functional requirements, and hardware and software requirements.

Chapter 4 contains system model and chapter 5 contain the proposed approach to solve imbalanced image set. Chapter 6 deal with system implementation and chapter 7 consists of unit and integrated testing of proposed approach. Chapter 8 entails with results and discussions and chapter 9 involves conclusion and future work.

**CHAPTER 2**

**LITERATURE SURVEY**

Writing survey is a rundown of the past work which is important to the theme. This section is fundamental advance to make the nature by considering the past research work. This section is a group of the Theoretical information and Methodologies identified with Imbalanced information.

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, et al., [9] describes what imbalance problem is, the factors affecting the datasets and methods to deal with the imbalanced data. There are two solution categories discussed- Sampling and cost-sensitive learning. There are two types of sampling- over-sampling and under sampling. The authors have compared both the sampling types. It is conveyed that over sampling outperforms under sampling while dealing with data-level approach.

Dr. D Ramyachitra, et al., [12] discusses the characteristics of imbalanced problem and the techniques to solve it. Authors describe five different levels to handle imbalance data- data level, algorithmic level, cost-sensitive level, feature selection level and ensemble level. Authors explained the methods and algorithms that come under each level to solve imbalance issue. The author discuss general steps to solve imbalance issue where authors under sample the majority class data and over sample the minority class data to balance the dataset. Authors describe on various databases like NCBI (National Centre for Bioinformatics Information), UNIPROT, UCI repository etc from where one can get the datasets. Various machine learning tools like WEKA, Keel, RapidMiner is explained and different evaluation metrics like ROC curve, precision, recall, cost matrix etc are explained which help in evaluating the performance of classifiers. Authors concluded that data-level approach-oversampling technique is best to solve the imbalance issue.

Hebbatallah Mostafa Anwer, et al., [7] reviewed the issues that one can attain while learning about the unfair datasets and the issues in class imbalance problem. The authors describe the approaches present to handle the issues, trends and advancements in unfair class categorization.

Tdd Perry, et al., [10] proposed an approach to automatically generate optimized cost matrix to solve categorization problems with any number of classes. They used genetic algorithm to get optimized cost matrix and compared their approach against un-optimized categorizers and different boosting algorithms by experimenting on various datasets. Authors conclude the performance of classifiers suffering from imbalance problem can be improved by cost matrix optimization.

Gregory Luppescu, et al., [5] explained the advance implemented- auto-enhancement skeleton which modifies images in a personalized way and on user preference. They discuss five components- Distance metric computation, to find training set which max represents dataset, to find optimal parameter set for training set, training it and Enhancing it. Validity of this method was tested and was concluded that this method is better than straight image enhancement.

Feature extraction play an important role when dealing with pictures as collection of pixels form a photo. Features provide the information to solve the imbalance issue and refer to structures from simple to complex types [8].

Ruchika Mishra, et al., [15] discuss the two approaches for synthetic sample creation-Spatial domain and frequency domain. The comparison between two approaches is made, advantages and disadvantages of both the approaches is discussed. It is hard to convey which approach is the best one to create synthetic samples of image.

Anu Namdeo, et al., [6] describe different synthetic image generation techniques for the better quality images and describe each method, their advantages and disadvantages.

Consideration of standard measures is important in order to evaluate the classifier performance, D. Druga Prasad, et al., [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. Authors experimented on 15 different datasets and convey the results. The WIMOTE method improve the performance measures well- accuracy, precision, recall, F-measure, AUC). In this project the specificity is measured. With the survey done, the proposed approach in improving performance is explained in further sections.

G N Srinivasa, et al., [20] provide overview of methods and algorithms for analysing structural texture of 2D images. Authors discussed what texture analysis is, the approaches like statistical and syntactic to analyse the texture of an image. Three principle statistical approaches- first order, second order and spectral in image processing was explained with their properties in detail. They conclude by describing what each approach does, spectral approach is best suited for periodic 2-D patterns in an image.

M AkkilJabbar, et al., [13] presented a novel approach to categorize the medical data authors combined KNN algorithm with genetic algorithm to come up with effective categorization. Around 7 datasets from UCI repository were considered and experiment on each dataset was performed. The proposed model improved the accuracy in diagnosis of heart diseases and helped doctors in diagnosing heart diseases with less attributes.

Wacharasak Siriseriwan, et al., [21] discuss class imbalance problem by SAFE level-SMOTE technique to generate synthetic images around original minority instances while avoiding the nearby majority images. The issue of losing few minor class data from this approach was focused and aimed to overcome it by combining two processes. The first is moving the minority images away from majority ones, the other one is to handle minor class images with I nearest neighbour model. The visible progress was shown in F-measure using various classifiers.

Yingying Qin, et al., [2] describe a method-3D modelling, which produce enough images to supplement original data in order to solve imbalance problem. Authors considered image set from internet and classified images into 7 categories (tank, helicopter, car, ship, fighter plane, fighting vehicle and armoured vehicle) and used 3d-max models to generate similar realistic images. Authors concluded that 3D max modelling helped in enlarging the dataset and helped in resolving imbalance issue.

**2.1 Summary:**

The Summarized Literature review explains about what is imbalanced data, methods to deal with imbalanced data in different areas. Authors describe the approaches to deal with imbalance issue and some future directions. Experiments are conducted on different real-time image sets using different classification algorithms. Authors discussed regarding the synthetic image generation techniques and the feature extraction part, they showcase that at data-level approach, over sampling is the best to solve imbalance issue.

**CHAPTER 3**

**REQUIREMENT ANALYSIS**

The chapter consist of the hardware and software requirements of this project and functional and non-functional requirements.

**3.1Requirement specification:**

**3.1.1 Functional requirements:**

The Functional necessities significantly grandstand the conduct of the Functionality and it's additionally fulfilling the information. In this undertaking the practical prerequisites are recorded underneath.

* **Image set collection:** Collection of fine images with similar properties for two categories is a first and necessary step. This helps in training the model to categorize the images.
* **Training the model:** Using any traditional classification algorithm, train the classifiers with the images collected of different categories.
* **Performance check:** Once the model is trained, it should be assessed to know how well the classifier did learn and this helps in improving the performance of the system and balance the image set.

**3.1.2 Non-functional requirements:**

* **Capacity:** The model works with any number of images and categories of data. The model estimates performance for considered image sets.
* **Availability:** The framework will be accessible in all conditions for all conditions with various choices.
* **Learnability:** The model learns from the features extracted and decides to which category the image falls under.
* **Reliability:** The Training model has to classify images correctly as expected result and the performance of the model is checked.
* **Adaptability:** The System conditions are able to adjust to any image collection and it demonstrates the genuine yield.

**3.2 Hardware and Software requirements:**

**3.2.1 Hardware requirements:**

* Processor : intel core ”i5”
* System Type : 64 bit/32 bit
* Ram : 4GB
* Hard disk : 500 GB

**3.2.2 Software Requirements:**

* Coding Language: MATLAB scripting
* Tool: MATLAB2017/MATLAB2013
* Operating System: Windows 7
* Image set : Animals (cat and dog), Forest and highway, Clean roads v/s garbage

**3.3 Summary:**

In this chapter, we discuss and specify the hardware and software requirements, functional and non-functional requirements for this project.

**CHAPTER 4**

**SYSTEM MODEL**

**4.1 Data Mining:**

Data mining is the process of examining large databases to generate new data, analysing different patterns of data with different perspective and converting it to useful information. It is also known as data discovery or knowledge discovery. It is broadly classified into three phases:

1. Prediction
2. Clustering
3. Categorization

As this project focuses on categorization, this chapter briefs on different categorization approaches and the learning approaches and methods.

**4.1.1 Categorization:**

It is the process of assigning a label to the existing classes. This is a two-step process- learning and categorization, learning phase is the one where the model is constructed and also involves training the model with a set of inputs. The constructed model is used to identify the set of labels for a given set of data and this is categorization phase.

* **Unlabelled and labelled data:**

Unlabelled data is the one that consist of natural samples or human created artefacts that can be obtained easily from the world. Examples include x-rays, audios, videos etc. Taking the set of unlabelled data and expanding each piece of it with a sort of meaning and labelling them forms a labelled data.

* **Cross validation:**

It is the way to estimate the performance of algorithm on new data, unknown data, given a set of data with known categories. It represents the (%) rate of correctness and incorrectness measures.

* **Bi-class categorization**:

It is the problem of categorizing the instances into one or two classes.

**4.1.2 Categorization approaches:**

* **Statistical:**

This approach involves humans, it will be characterized by precise fundamental probability models i.e., it will estimate the probability of being in each class instead of categorization.

* **Neural networks:**

This approach consist of layers of interconnected nodes where each node produce a non-linear function of the input. The input to a node may come directly from input data or from the other node. This approach involves patterns and decision making, the nodes are identified with the network output.

* **Machine learning:**

It is the approach that involves auto-computations or the procedures based on logical or binary operations that can learn tasks from set of examples.

**4.1.3 Machine learning approaches:**

* **Data Level:**

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:**

This approach deal with modifying existing learning algorithms to ease the favoritism 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 project concentrates on data-level approach.

**4.2 Learning methods:**

**4.2.1 Supervised:**

Analysing the data used for training and producing inferred function which can be used for mapping new scenarios or inferring the function from labelled data is called supervised learning. This is mainly related to machine learning and deals with classification of data. The problems include classification and regression.

**4.2.2 Semi-supervised:**

It is a class of supervised learning tasks that make use of unlabelled data for training and perform analysis. The input to this learning is a mixture of labelled and unlabelled data, the problems include classification and regression.

**4.2.3 Unsupervised:**

This draws inferences from data consisting of input data without labelled responses. Example: cluster analysis. The problems include clustering, dimensionality reduction and association rule learning.

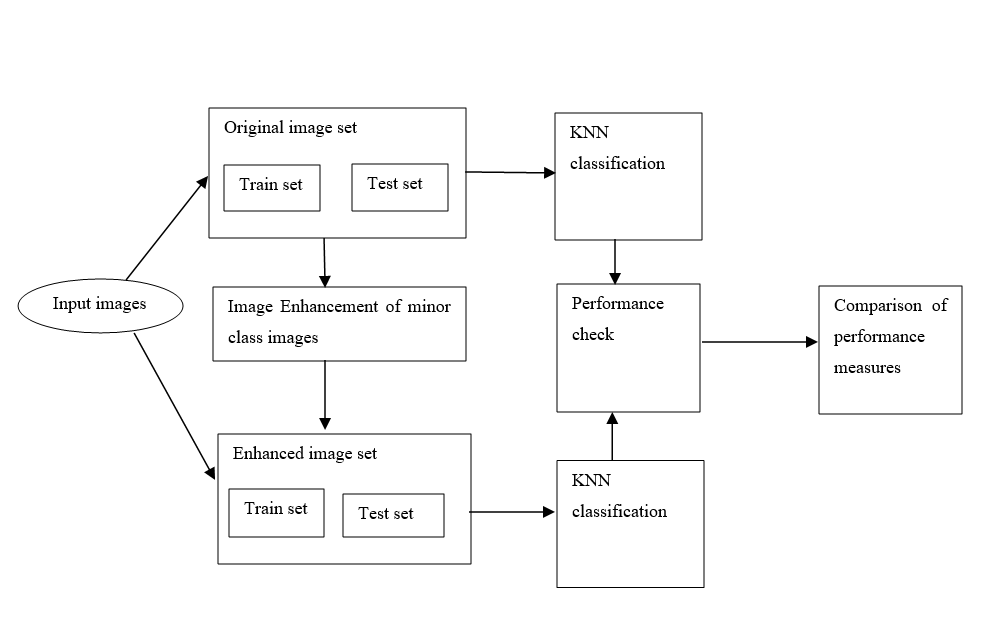
**4.3 Summary:**

This chapter discussed what data science is all about, the three phases of it. It mainly focuses on categorization and its approaches, learning method, machine learning and its approaches and convey that data-level approach is followed in this project.

**CHAPTER 5**

**PROPOSED SYSTEM DESIGN AND ANALYSIS**

**5.1 System Architecture:**

****

**Figure 1: Proposed system architecture to handle Imbalanced Image set**

In this approach, we provide different bi-class image sets as inputs and consider some images for training and some for testing. First stage of project deals with applying KNN classification on train set, train the model, 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 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 enhanced 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 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.

**5.1.1 Original Set:**

The Original set consist of two folders- train and test, we train the categorizer from the original set initially and check the performance of the system considering pictures from test set.

**5.1.2 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

(1)

where ‘p’ represent pixel value at point i, j for MxN sized image [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

(2)

which represent root of pixel ‘p’ at points i, j for MxN image 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= (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 miss-categorization. It is a distribution variation in an area. It can be represented as [8][19]

(4)

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

**5.1.3 KNN Classification:**

Features of both train set pictures and test set pictures are extracted and stored in a .mat file. The extracted features are compared and based on this comparison the performance measures is estimated. In this project, K-Nearest Neighbour is considered for categorization of images, where the k value assigned is 5.

**5.1.4 Modified Set:**

Considering the same set of images that are in original set, synthetic image generation techniques discussed above are applied on the minor class images and stored in modified set. Now, categorization is performed on the modified set and the performance measures are estimated to check if there is any improvement in the measures.

**5.1.5 Performance Check 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 samples identified as negative out of total number of negatives. It is also known as True negative rate and is assessed as:

Specificity= True Negative/ (True Negative + False Positive) (5)

* ***Recall:***

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

Recall = True Positive/ (True Positive+ False Negative) (6)

* ***Precision:***

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

Precision = True Positive / (True Positive + False Positive) (7)

**5.1.6 Synthetic Image Generation 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 attempts to progress an image by elongating the range of intensity values it contains in order to make full use of possible values. Contrast stretching is constrained to a linear 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 scale, so all the picture detail is compacted towards the dark tip of the histogram. If we could ‘expanse out’ the grey levels at dark tip to produce a more uniform dispersed histogram then the picture would get 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:***

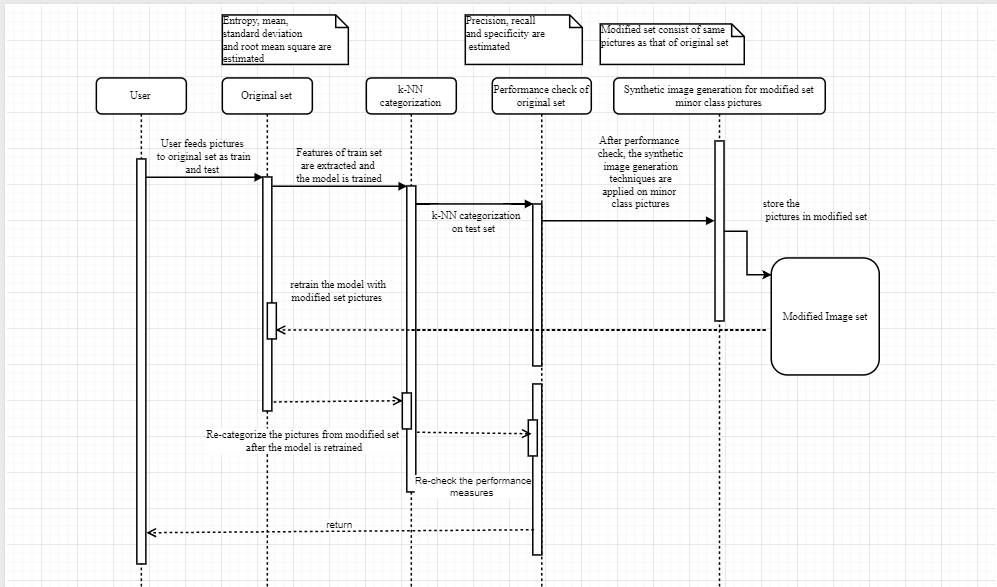
As brightness preserving is an important feature of an image, this method is used to shield the illumination of an image by splitting the picture’s histogram into two equal parts in order to make sure that the strengths are also equally arranged [6].

* ***Adaptive histogram equalization:***

Versatile HE is used for enhancing contrast as a piece of pictures. It fluctuates from Histogram Equalization by adaptable procedure that figures a couple of histograms and each histogram identifying with a specific fragment of a picture. The distinction of zone for a picture won't be satisfactorily enhanced by Histogram Equalization. The method helps this advancement by adjusting each pixel with a change limit got from a region district. It is used to vanquish a couple of checks of overall direct min-max windowing system. Thus it diminishes the proportion of bustle in areas of an image. Besides, this technique have the limit with respect to upgrading the distinction of grayscale and shading picture [6].

**5.2 Sequence diagram:**

Sequence diagram is the one which deal with the process and interaction between the phases or modules, it explains the project flow and the scenarios of where imbalance issue occurs. The diagram below represents UML (Unified Modelling Language) for this project.

****

**Figure 2 Sequence diagram for proposed approach**

The figure explains the project flow, the user trains the model by feeding pictures of two categories in folder named ‘Original set', the folder contains sub-folders of two different classes. Features of every picture from original set including the both classes is extracted and stored, user performs KNN categorization on test images and estimates the performance of the system. The features extracted and the measures computed are discussed earlier. Considering same set of images in modified set, synthetic picture generation techniques are applied to the minor class pictures of modified set, user retrains the system with generated pictures along with the original pictures, the steps repeat and again the performance measures are estimated to check for the improvement if any.

**5.3 Use case diagram:**

Use case diagram represents the real world scenarios with respect to this project. It explains the set of actions that take place in the entire project. The following diagram depicts the use case diagram for imbalanced image set.



**Figure 3 Use case diagram of proposed approach**

* **Input pictures:**

Different category pictures are given as input to the system, the images is of two categories (bi-class), the image set considered are cats v/s dogs, clean roads v/s garbage and forest v/s highway.

* **Feature Extraction:**

The features of each image is calculated and stored in a file (.MAT), the features considered are discussed above. Features of both the class is computed and the test set features also calculated.

* **KNN Categorization:**

User performs k-nearest neighbour categorization on the test images based on the pictures the system is trained, the k-nn categorizer categorizes the test set pictures in one of the two classes considering the features extracted. The test set consist pictures of both the classes and the value of k is set to 7 in this experiment.

* **Performance check:**

After the categorization of test set, the measures- precision, recall, specificity of the classifier is estimated.

* **Synthetic picture generation:**

User creates synthetic pictures for all the minor class pictures of modified set and stores it back in the same folder along with the original pictures. The synthetic pictures are generated by using the techniques discussed above.

* **Retrain the model:**

The user retrains the model with modified image set, performs k-nn categorization on the same in order to recheck the performance of the system. The system is retrained and tested with every synthetic picture generation method i.e, each method is applied on every picture of minor class, the test is performed with iteration of methods.

* **Performance recheck:**

Once the categorizer is retrained with modified image set, the performance is recalculated in order to check if any improvement in the measures with respect to all the methods.

**5.4 Summary:**

Proposed architecture explains sequence diagram and Use Case diagram. The system architecture is composed of four major modules- Training the model, categorizing the images, checking the performance of system, creating synthetic images of minor class pictures (modified set) and rechecking the performance measures. Sequence diagram which describes the order in which they occur. Use case diagram discuss about the real time scenarios of the system. The same flow is continued with all the synthetic image generation techniques specified above.

**CHAPTER 6**

**SYSTEM IMPLEMENTATION**

System implementation is the process of converting the designed system architecture into working modules where it is made sure that all the functional and non-functional requirements are met.

**6.1 Algorithm for Classification:**

K nearest neighbour is a supervised classification algorithm which takes a bunch of labelled points and use them to learn how to label other points. To label a new point, it will look at the labelled points that are closest to the new point. Based on the neighbour votes, whichever label most of the neighbours have is the label assigned to the new point. The number of neighbours that it checks is based on’ k’ value. In this project Euclidian distance is measured and the k value considered is 5.

***Alogorithm:***

**classify (X, Y, x)** // X-training data, Y->class labels, x->unknown samples

**for i= 1 to m do**

**compute distance d (Xi, x)**

**end for**

**compute set I containing indices for k smallest distance d (Xi, x)**

**return majority label for {Yi, where iI}**

***Euclidian distance*: (5)**

**6.2 Algorithms of Synthetic Image generation techniques:**

**Histogram Equalization:**

The histogram of an image is nothing but the discrete function formed by counting the number of grey pixels in the image. When this function is normalized to sum up to 1 for all grey level values, it is treated as “Probability density function”. It shows how probable is certain grey value found in a image. The algorithm to calculate the histogram of an image is given below.

***Algorithm:***

Input: Image X

Output: Histogram of image Y

begin

[r, c] = size(X);

for i = 1: r //r is a row and c is a column

for j = 1: c

hist{x[i,j]} = hist{x[i, j]}+1

end for

end for

for g = 1 : Gmax

hist[g]=hist[g] / (M\*N)

end for

end

***Probability density function***:

****

**Figure 4: Input image before Histogram equalization**

****

**Figure 5: Output image after Histogram equalization**

**Adaptive Histogram Equalization:**

This method focus on enhancing the contrast of local region rather than the entire image. Each pixel is ranked by its intensity level comparative to its neighbouring pixel’s intensity values. After the ranking process, the pixel is assigned a new value from the available intensity range to its rank. This new value is assigned to output image so as to not disturb the ranking of original pixels.

***Algorithm:***

for each (x,y) in image do

{

rank = 0

for each (i, j) in contextual region of (x, y) do

{

if image[x,y] > image[i,j] then

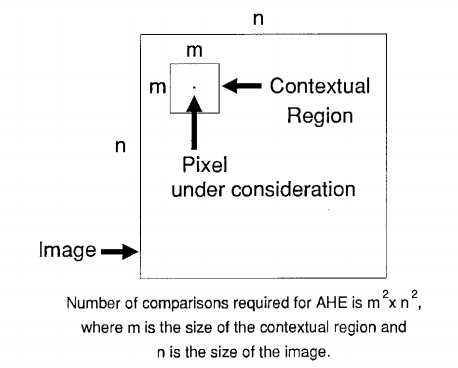
rank = rank +1

}

output[x,y] = rank \* max\_intensity / (# of pixels in contextual region)

}

Contextual region:



**Figure 7 : Contextual region**

****

**Figure 8: Input image before adaptive hist equalization**

****

**Figure 9: Output image after adaptive hist equalization**

**Contrast Stretching:**

Contrast of an image is nothing but the spread of its histogram. The dynamic range of an image is the entire range of the intensity values in an image, in simple terms, maximum pixel value minus minimum pixel value. Increasing the dynamic range lead up to stretching the histogram, so basically this algorithm pulls the boundaries of original histogram function to extreme.

***Algorithm:***

Input : q bit image // 8 bit image

MP // eg: 255 for 8-bpp image, MP: maximum pixel

a= min (i)

b= max (i)

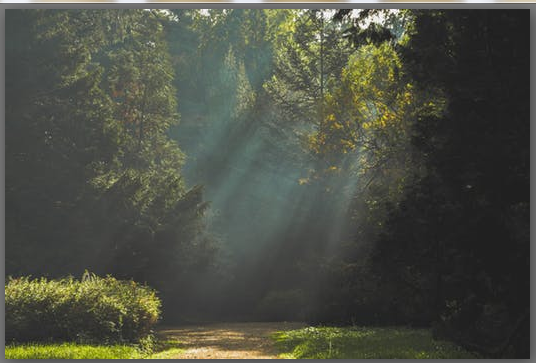
R = b-a

for each (pixel p in I)

p= [(p-a)/R]MP // applies linear scaling function

p =round(p)

end



**Figure 10: Input image before contrast stretch**



**Figure 11: Output image after contrast stretch**

**Contrast enhancement:**

**Brightness Preserving Bi-histogram equalization:**

**6.3 MATLAB tool:**

MATLAB (Matrix laboratory) is a case sensitive language developed by Mathworks which works with matrices i.e, a variable c is very different from C. Everything the MATLAB understands is in a form of matrix, it is a high-level language which consist of various toolboxes which makes things much easier. The MATLAB code is easy to learn and is optimized to be quick when dealing with matrix operations. MATLAB is procedural language and possess some object-oriented elements. MATLAB can behave like a calculator or as a programming language, the errors here are easier to fix. Several matrix operations, mathematical operations and image related operations, several algorithms can be run in MATLAB with a single command. Every year Mathworks release the new toolboxes and upgrade the version of software.

In this project, we install MATLAB 2017 in our PC and work on the same, MATLAB is widely used in engineering streams which includes audio and video processing, control systems, test and measurements, signal processing and communications etc.

**6.4 Summary:**

The chapter explains the k-nn algorithm used for categorization of images, the algorithms of the image enhancement techniques and the output of each technique after applying them on original image.

**CHAPTER 7**

**TESTING**

Testing is one of the way of assessing the system which helps to detect the quality of the software, the methods we follow and evaluates the expected output and actual input. Verification and Validation are the process in the software testing where we verify various things and validate them. Some of the conditions are stated at the development phase which must be satisfied by the product is called the Verification. The Requirement must be specified at the end of the development phase which assures the Validation.

**7.1 Unit Testing**

The initial testing which deal with testing of individual modules of the system is known as Unit testing or white box testing. The Below Table shows the Test Case modules drawn out to make sure everything is tested at its best.

|  |  |
| --- | --- |
| Testcase\_ID | UTC01 |
| Name | Loading the image set |
| Description | Load the dataset to original set of both classes |
| Input | Pictures of both categories in original set |
| Expected output | The Dataset Loaded successfully with two class categories images |
| Actual Result | Same as expected output |
| Status | Pass |

**Table 1: Unit Test Case 1 - To load the image set in a folder**

|  |  |
| --- | --- |
| Testcase\_ID | UTC02 |
| Name | Select the image directory path |
| Description | Specify the path name in code to choose the directory |
| Input | Path of the image folder (Original/ modified) |
| Expected output | Path accessed and pictures selected |
| Actual Result | Same as expected output |
| Status | Pass |

**Table 2: Unit Test Case 2- To select the path of images stored**

|  |  |
| --- | --- |
| Testcase\_ID | UTC03 |
| Name | Picture feature extraction |
| Description | Calculate the mean, standard deviation, entropy and root mean square of all pictures in train and test set |
| Input | Pictures in original train and test set |
| Expected output | .MAT file with all features extracted |
| Actual Result | Same as expected output |
| Status | Pass |

**Table 3: Unit Test Case 3- Extract features of images in train set**

|  |  |
| --- | --- |
| Testcase\_ID | UTC04 |
| Name | Verify performing k-NN categorization on test set |
| Description | Run k-NN on test set pictures |
| Input | K value, k-NN function |
| Expected output | All the pictures categorized in either category/class |
| Actual Result | Same as expected output |
| Status | Pass |

**Table 4: Unit Test Case 4- Perform k-NN on test set**

|  |  |
| --- | --- |
| Testcase\_ID | UTC05 |
| Name | Verify the performance measures |
| Description | After k-NN categorization, estimate performance measures (precision, recall, specificity) |
| Input | k-NN categorization |
| Expected output | Precision, recall and specificity measures displayed in UI |
| Actual Result | Same as expected output |
| Status | Pass |

**Table 5: Unit Test Case 5- To verify the performance measures**

|  |  |
| --- | --- |
| Testcase\_ID | UTC06 |
| Name | Verify synthetic picture generation |
| Description | To create synthetic picture generation of minor class pictures from modified train set |
| Input | Pictures in modified set |
| Expected output | Pictures generated in modified image set with respected technique applied |
| Actual Result | Same as expected output |
| Status | Pass |

**Table 6: Unit Test Case 6- To verify creating synthetic pictures of minor class**

|  |  |
| --- | --- |
| Testcase\_ID | UTC07 |
| Name | Verify to retrain the model with modified set |
| Description | Re-training the model with modified set |
| Input | Original pictures with synthetic generated pictures in modified set |
| Expected output | The model retrained |
| Actual Result | Same as expected output |
| Status | Pass |

**Table 7: Unit Test Case 7- Retrain the model**

|  |  |
| --- | --- |
| Testcase\_ID | UTC08 |
| Name | Verify to run k-NN on modified set |
| Description | After retraining the model, to categorize the pictures |
| Input | Features of retrained system |
| Expected output | Features extracted and retrained system categorized |
| Actual Result | Retrained system categorized |
| Status | Pass |

**Table 8: Unit Test Case 8- Re-categorize the modified set**

|  |  |
| --- | --- |
| Testcase\_ID | UTC09 |
| Name | Verify to check the performance of modified set |
| Description | Validation |
| Input | k-NN classification and features extracted |
| Expected output | Improvement in performance measures |
| Actual Result | Same as Expected |
| Status | Pass |

**Table 9: Unit Test Case- Verify to check the performance of modified set**

**7.2 Integration Testing**

Integration Testing is the one which clubs all the test cases and constitute one sub system to make sure all modules are working properly and produce the expected Result.

Integration Test Case are listed below

|  |  |
| --- | --- |
| Testcase\_ID | ITC10 |
| Name | Integration Test cases of training the model, categorizing, performance checking |
| Description | Training, categorization and performance check |
| Input | Pictures in original set of different categories |
| Expected output | k-NN categorization performed and performance measures displayed |
| Actual Result | Same as Expected |
| Status | Pass |

**Table 10: Integration Test case 1 – Integration testing of the entire mentioned algorithm and performance check for original set**

|  |  |
| --- | --- |
| Testcase\_ID | ITC11 |
| Name | Integration of all test cases- synthetic picture generation, retraining the model, performance check of modified set |
| Description | Synthetic picture generation, retraining the model, re-categorization and check for performance improvement |
| Input | Synthetic pictures generated in modified set along with original set pictures |
| Expected output | System retrained, modified set test pictures re-categorized, performance improvement |
| Actual Result | Same as Expected |
| Status | Pass |

**Table 11: Integration Test case 2 – Integration testing of the entire mentioned algorithm in distributed IDE**

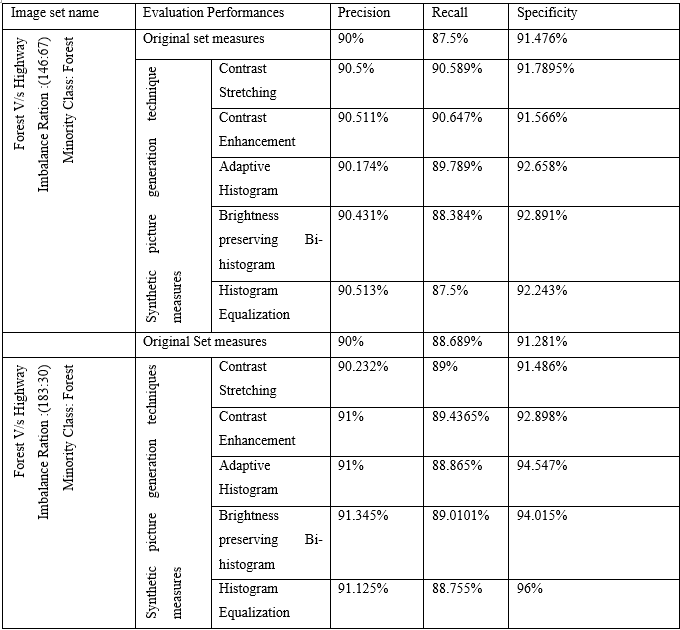
**7.3 Summary:**

The Chapter includes all the Unit Test Case and Integration Test Cases which efficiently check the correctness of data. Here all the test cases are working as expected and the modules worked as expected.

**CHAPTER 8**

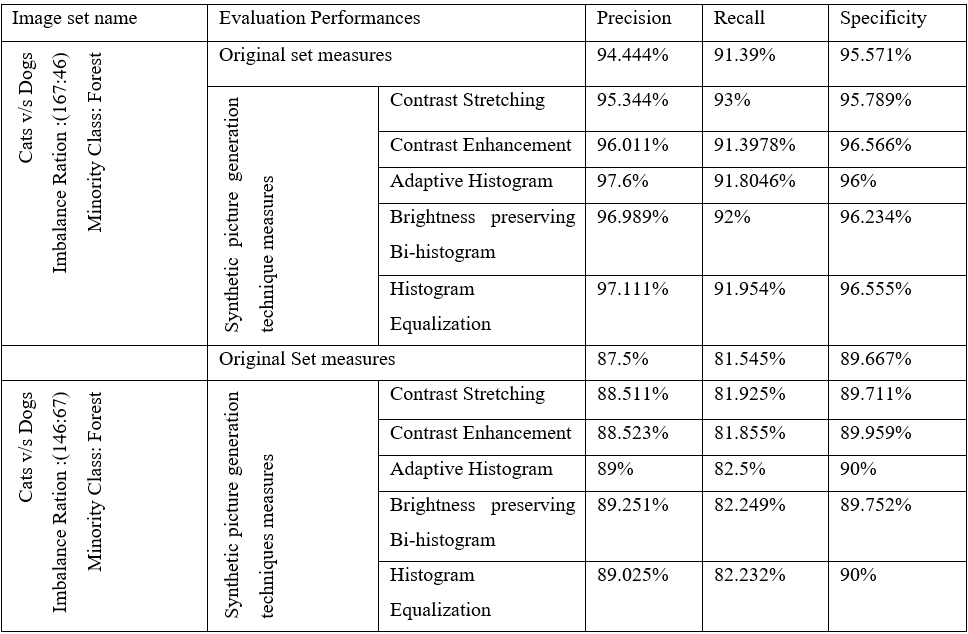
**RESULT AND DISCUSSIONS**

***Image set description:*** Image set 1- The first image set considered is the forest and highway where the major class is the highway class and the minor class is forest. The images are considered from the [www.pexels.com](http://www.pexels.com) website with similar properties.



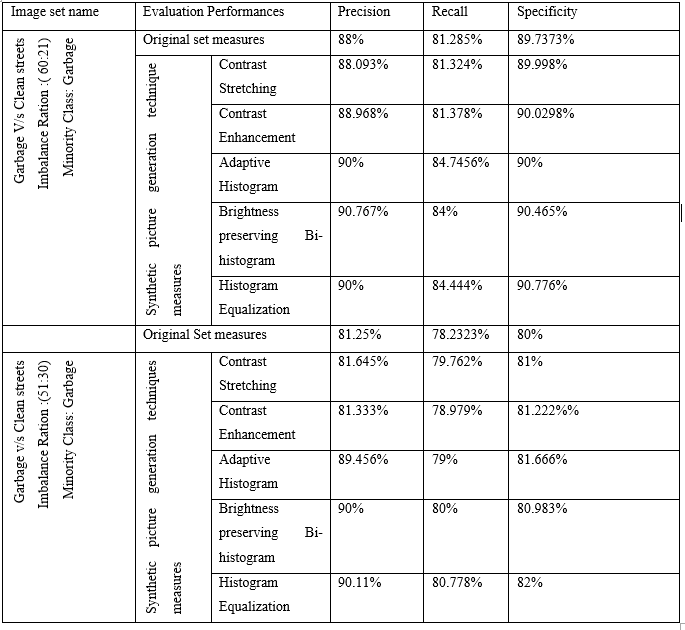
**Table 12: Synthetic sample generation for Forest v/s Highways image set**

***Image set description:*** Image set 2- The second image set considered is cats v’s dogs where the major class is cat’s class and the minor class is dog’s. The images are considered from the [www.imagenet.com](http://www.imagenet.com) website with similar properties.



**Table 13: Synthetic sample generation for cats v/s dogs image set**

***Image set description:*** Image set 3- The third image set considered is garbage v/s clean streets where the major class is clean street class and the minor class is garbage. The images are considered from the [www.pexels.com](http://www.pexels.com) website with similar properties.



**Table 14: 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 with 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, improvements were visible in all techniques and with the ratio- (146:67) improvement in recall and specificity were visible 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 progress with all techniques discussed. By this approach we infer that using above techniques on minor class images helps in progress of the system performance.

**CHAPTER 9**

**CONCLUSION AND FUTURE WORK**

The project aims to limit the bias towards majority class data for an imbalanced image set. The proposed approach uses KNN classifier to categorize the image set and the performance is evaluated using different evaluation measures. Synthetic samples of minor class images are created using synthetic image generation techniques and the system is retrained with synthetic and original images. The experiment shows improvement in the measures-precision, recall, specificity when the performance is re-calculated.

The future work focus on dealing with more intricate image sets and apply different synthetic image generation techniques. The various techniques like contrast limited adaptive histogram equalization; which helps in improving every small area of image by reducing noise present in the image. To present high frequency content images and to improve over dark images, adaptive DWT based DSR technique can be used. Adaptive speckle filter technique help in improving the quality of blurred images, this method make use of dynamic brightness assignment to escalate intensity value of image that help in brightening of image.

Non-Linear filtering algorithms like mean filter, vector median filter and modified spatial median filter; can be used to smoothen the image by calculating the average number of pixels affected by noise and removing the noise present in the image. Vector median filter fills the purpose of removing noise and fines the image. Other filtering technique like homomorphic filtering handle the lightening conditions and sharpens the image. A combination of two or more techniques specified above can also be used in future to improve the image quality and generate many images to solve imbalance issue.

**REFERENCES**

1. Dr. Durga Prasad, Dr. K Nageswar Rao, “An Improved 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: Mdelling 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*” 2017 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 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.