Textural Feature Extraction of Natural Objects for Image Classification

Vishal Krishna *vkrishna7@gatech.edu*
Computer Science
Georgia Institute of Technology
Atlanta – 30332, US

Ayush Kumar *f2010029@goa.bits-pilani.ac.in*
Computer science
BITS Pilani, Goa Campus
Goa – 403726, India

Kishore Bhamidipadi *Kishore.b@manipal.edu*
Computer Science Engineering
Manipal Institute of Technology
Manipal – 576104, India

Abstract

The field of digital image processing has been growing in scope in the recent years. A digital image is represented as a two-dimensional array of pixels, where each pixel has the intensity and location information. Analysis of digital images involves extraction of meaningful information from them, based on certain requirements. Digital Image Analysis requires the extraction of features, transforms the data in the high-dimensional space to a space of fewer dimensions. Feature vectors are n-dimensional vectors of numerical features used to represent an object. We have used Haralick features to classify various images using different classification algorithms like Support Vector Machines (SVM), Logistic Classifier, Random Forests Multi Layer Perception and Naïve Bayes Classifier. Then we used cross validation to assess how well a classifier works for a generalized data set, as compared to the classifications obtained during training.

.**Keywords:** Feature Extraction, Haralick, Classifiers, Cross-Validation.

1. INTRODUCTION

Texture is an important feature for many types of analysis of images and identification of regions of interest. Texture analysis has a wide array of applications, including industrial and biomedical monitoring, classification and segmentation of satellite or aerial photos, identification of ground relief, and many others. [1] Various methods have been proposed via research over the years for identifying and discriminating the textures. Measures like angular second moment, contrast, mean, correlation, entropy, inverse difference moment, etc. have been typically used by researchers for obtaining feature vectors, which are then manipulated to obtain textural features. One of the most popular approaches to texture analysis is based on the co-occurrence matrix obtained from images, proposed by Robert M. Haralick in 1973, which forms the basis of this paper.

Image classification is one of the most important part of digital image analysis. Classification is a computational procedure that sorts images into subsets according to their similarities. [4] Contextual image classification, as the name suggests, is a method of classification based on the contextual information in images, i.e. the relationship amongst neighbouring pixels. [2].

For classification, we used the WEKA ("Waikato Environment for Knowledge Analysis") tool, which is an open source machine-learning software suite developed using Java, by the University of Waikato, New Zealand.[6] It contains set of tools for different data analysis and modelling techniques such as: pre-processing, classification, clustering, segmentation, association rules and visualization. It implements many artificial intelligence algorithms like decision trees, neural networks, Particle Swarm Optimization etc.).[5]

2. LITERATURE SURVEY

The classification of images can be done either on the basis of a single resolution cell or on a collection of resolution cells. When a block of cells are used, the challenge is to define a set of features to represent the information given by those cells, which can be used for classification of the images.

Human perception of images is based on three major classes of features: spectral, textural and contextual. Spectral features are obtained as the average variation of tone across various bands of the electromagnetic spectrum. Textural features, on the other hand, provide information about the variation of tone within a single band. Information from portions of image surrounding the part under analysis constitute the contextual features. In gray-scale photographs, tone represents the varying gray levels in resolution cells, while the statistical distribution of the gray levels is interpreted as texture. Tone and texture form an intrinsic part of any image, though one can get precedence over the other according to the nature of the image. Simply stated, the relation between the two is: tone is dominant when the sample under consideration shows only small range of variation of gray levels, while gray levels spread over a wide range in a similar sample indicate the dominance of texture.

Haralick’s work is based on the assumption that information regarding the texture of any image can be obtained from calculating the average spatial relation of the gray tones of the image with each other. The procedure for calculating the Haralick textural features is based on a set of gray-tone spatial-dependence probability distribution matrices (also termed as Gray-Level Co-occurrence Matrices or GLCM, or gray-level spatial dependence matrix), computed for various angles at fixed distances. From each such matrix, fourteen features can be calculated, which provide information in terms of homogeneity, contrast, linear variation of gray tone, nature and number of boundaries etc.

Co-occurrence Matrix: A co-occurrence matrix, P, is used to describe the relationships between neighbouring (at a distance, d) pixels in an image. 4 co-occurrence matrices, each calculated for a different angle, can be defined. A co-occurrence matrix, termed as P^0, describes pixels that are adjacent to one another horizontally (at angle 0^o). Similarly, co-occurrence matrices are defined for the vertical direction (90^o) and both diagonals (45^o and 135^o). These matrices are called P^{90}, p^{45} and P^{135} respectively. [3]

$$x = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 2 & 2 & 2 \\ 2 & 2 & 3 & 3 \end{pmatrix} \quad p_0 = \begin{pmatrix} 4 & 2 & 1 & 0 \\ 2 & 4 & 0 & 0 \\ 1 & 0 & 6 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$$

$$p_{45} = \begin{pmatrix} 4 & 1 & 0 & 0 \\ 1 & 2 & 2 & 0 \\ 0 & 2 & 4 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} p_{90} = \begin{pmatrix} 6 & 0 & 2 & 0 \\ 0 & 4 & 2 & 0 \\ 2 & 2 & 2 & 2 \\ 0 & 0 & 2 & 0 \end{pmatrix}$$

$$p_{135} = \begin{pmatrix} 2 & 1 & 3 & 0 \\ 1 & 2 & 1 & 0 \\ 3 & 1 & 0 & 2 \\ 0 & 0 & 2 & 0 \end{pmatrix}$$

There are 4 pairs of (0,0) in angular 0, thus $P^0(0,0)=4$, there are 2 pairs of (0,1), thus P0(0,1)=2. Similarly all the four matrices are computed.

Based on the co-occurrence matrices calculated as above, the thirteen texture features as proposed by Haralick are defined below:

Notation:
N_g : Number of distinct gray levels in quantized image

$$p_x(i) = \sum_{j=1}^{N_g} p(i,j)$$

$$p_y(j) = \sum_{i=1}^{N_g} p(i,j)$$

$$p_{x-y}(k) = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j),\ |i-j| = k \text{ and } k = 0,1,\ldots,N_g-1$$

$$p_{x+y}(k) = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j),\ i+j = k \text{ and } k = 2,3,\ldots,2N_g$$

a) Angular Second Moment

$$f_1 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)^2$$

b) Contrast

$$f_2 = \sum_{k=0}^{N_g-1} k^2 p_{x-y}(k)$$

c) Correlation

$$f_3 = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(ij)p(i,j) - \mu_x\mu_y}{\sigma_x\sigma_y}$$

Where μ_x, μ_y, σ_x, σ_y are mean of x, y and standard deviation of x, y respectively.

d) Sum of Squares: Variance

$$f_4 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-\mu)^2 p(i,j)$$

e) Inverse Difference Moment

$$f_5 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}\frac{1}{1+(i-j)^2}p(i,j)$$

f) Sum Average

$$f_6 = \sum_{i=2}^{2N_g} i p_{x+y}(i)$$

g) Sum Variance

$$f_7 = \sum_{i=2}^{2N_g} (i - f_8)^2 p_{x+y}(i)$$

h) Sum Entropy

$$f_8 = -\sum_{i=2}^{2N_g} p_{x+y}(i) \log \left(p_{x+y}(i)\right)$$

i) Entropy

$$f_9 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \log \left(p(i,j)\right)$$

j) Difference Variance

$$f_{10} = \text{variance of } p_{x-y}$$

k) Difference Entropy

$$f_{11} = -\sum_{i=0}^{N_g - 1} p_{x-y}(i) \log \left(p_{x-y}(i)\right)$$

l) Information measures of correlation

$$f_{12} = \frac{f_9 - HXY1}{\max(HX, HY)}$$

$$f_{13} = \sqrt{1 - \exp^{-2(HXY2 - f_9)}}$$

where HX and HY are entropies of p_x and p_y.

3. METHODOLOGY

For any value of d, as mentioned before, 4 matrices are calculated for each of the thirteen features detailed above. The mean and range of each set of four values give a set 28 values which are then passed to the classifier. Out of the input features, some share a strong correlation, so a feature-selection procedure can identify a subset of features in order to give good results in classification.

The test data has a total of 25 classes, which are known Apriori. We use this knowledge to calculate the effectiveness of various classification algorithms available, on the Haralick features. The classification algorithms used are:

1. Naïve Bayes Classifier (NB) - A Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. [7]
2. Logistic Classifier (Log) - Logistic regression is a probabilistic statistical classification model. It measures the relationship between the categorical dependent variable and one or more independent variables, which are usually (but not necessarily) continuous, by using probability scores as the predicted values of the dependent variable.[8]

$$HXY1 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \log \left(p_x(i) p_y(j)\right)$$

$$HXY2 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_x(i) p_y(j) \log \left(p_x(i) p_y(j)\right)$$

3. Multilayer Perception Classifier (MP) – In conventional MLP, components of feature vectors are made to take crisp binary values, and the pattern is classified according to highest activation reached. [9]

4. Random Forest Classifier (RF) - Random forests operate by constructing a number of decision trees training data and classifying data according to the mode of the obtained.[10]
5. Sequential Minimal Optimization – The algorithm is used to train support vector machines for classification. [11]

The parameters on which the effectiveness of each of the above algorithms are:

1. True Positive Rate (TP) – it is the number of items correctly labelled as belonging to the particular class divided by the total number of elements labelled as belonging to that class
2. False Positive Rate (FP) – it is the number of items incorrectly labelled as belonging to the particular class divided by the total number of elements labelled as belonging to that class
3. Precision - it is the fraction of retrieved instances that are relevant
4. Recall - it is the fraction of relevant instances that are retrieved
5. F-Measure – it is a measure that combines precision and recall, calculated as the harmonic mean of precision and recall
6. ROC Area - receiver operating characteristic (ROC) is a plot of the performance of a binary classifier system. The area under the curve is treated as a measure of accuracy of the classifier.

A second set of experiments are carried out, using the same test data, algorithms and parameters, but with the added constraint of using cross validation factor of 10.

4. RESULTS AND ANALYSIS

Each algorithm is first run on the data set and all six parameters are measured and compared. The results obtained are given below.

Class	TP Rate				
	NB	Log	MP	SMO	RF
1	0.525	1.000	0.950	0.675	1
2	0.750	1.000	0.975	0.675	1
3	0.850	1.000	1.000	0.875	0.975
4	0.775	0.975	0.975	0.800	1
5	0.900	1.000	1.000	0.975	1
6	0.825	1.000	0.975	0.800	1
7	0.900	1.000	1.000	0.950	1
8	0.775	1.000	0.925	0.700	1
9	0.850	0.975	0.825	0.675	1
10	0.800	1.000	1.000	0.925	1
11	0.725	1.000	0.975	0.775	1
12	0.750	1.000	0.950	0.825	1
13	0.800	1.000	1.000	0.900	1
14	0.650	0.975	0.975	0.850	1
15	0.725	0.975	1.000	0.850	0.975
16	0.850	1.000	0.975	0.800	0.975
17	0.975	0.975	1.000	0.800	1
18	0.900	0.975	1.000	0.975	1
19	0.600	0.975	0.975	0.800	1
20	1.000	1.000	1.000	1.000	1
21	0.250	1.000	0.925	0.725	0.975
22	0.750	1.000	0.975	0.900	1
23	0.525	1.000	0.950	0.775	0.9

24	1.000	1.000	1.000	0.975	1
25	0.850	1.000	1.000	0.975	1

FIGURE 1.1: Values of TP Rate of each class for different classification methods.

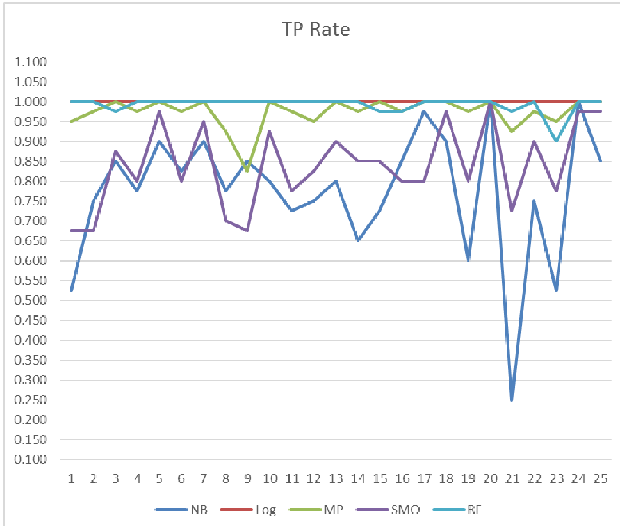

FIGURE 1.2: Graphical representation of TP Rate values.

Class	NB	Log	MP	SMO	RF
1	0.013	0.000	0.000	0.01	0.001
2	0.019	0.000	0.002	0.01	0
3	0.000	0.000	0.001	0	0.001
4	0.013	0.001	0.003	0.01	0
5	0.010	0.000	0.000	0	0.002
6	0.006	0.000	0.003	0	0
7	0.002	0.000	0.000	0	0
8	0.004	0.000	0.001	0.01	0
9	0.054	0.001	0.001	0.02	0
10	0.006	0.000	0.002	0.01	0.002
11	0.007	0.000	0.004	0.01	0.001
12	0.013	0.000	0.001	0.01	0
13	0.003	0.000	0.001	0	0
14	0.015	0.000	0.000	0.02	0
15	0.011	0.000	0.001	0	0.001
16	0.018	0.000	0.000	0	0
17	0.002	0.001	0.001	0.01	0
18	0.000	0.001	0.000	0	0
19	0.004	0.000	0.002	0.02	0
20	0.000	0.000	0.000	0	0
21	0.017	0.000	0.003	0.02	0
22	0.013	0.000	0.000	0.01	0
23	0.008	0.000	0.001	0.01	0
24	0.000	0.000	0.000	0	0
25	0.000	0.000	0.000	0	0

FIGURE 2.1: Values of FP Rate of each class for different classification methods.

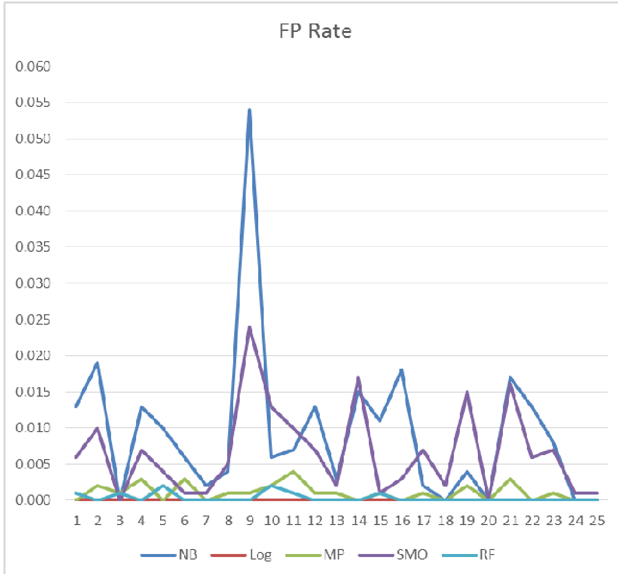

FIGURE 2.2: Graphical representation of FP Rate values.

Class	NB	Log	MP	SMO	RF
1	0.636	1.000	1.000	0.82	0.976
2	0.625	1.000	0.951	0.73	1
3	1.000	1.000	0.976	1	0.975
4	0.721	0.974	0.929	0.82	1
5	0.783	1.000	1.000	0.91	0.952
6	0.846	1.000	0.929	0.97	1
7	0.947	1.000	1.000	0.97	1
8	0.886	1.000	0.974	0.85	1
9	0.395	0,983	0.971	0.54	1
10	0.842	1.000	0.952	0.76	0.952
11	0.806	1.000	0.907	0.76	0.976
12	0.714	1.000	0.974	0.83	1
13	0.914	1.000	0.976	0.95	1
14	0.650	0.994	1.000	0.68	1
15	0.725	0.992	0.976	0.97	0.975
16	0.667	1.000	1.000	0.91	1
17	0.951	0.978	0.976	0.82	1
18	1.000	0.984	1.000	0.95	1
19	0.857	0.993	0.951	0.7	1
20	1.000	1.000	1.000	1	1
21	0.385	1.000	0.925	0.66	1
22	0.714	1.000	1.000	0.86	1
23	0.724	1.000	0.974	0.82	1
24	1.000	1.000	1.000	0.98	1

25	1.000	1.000	1.000	0.98	1

FIGURE 3.1: Values of Precision of each class for different classification methods.

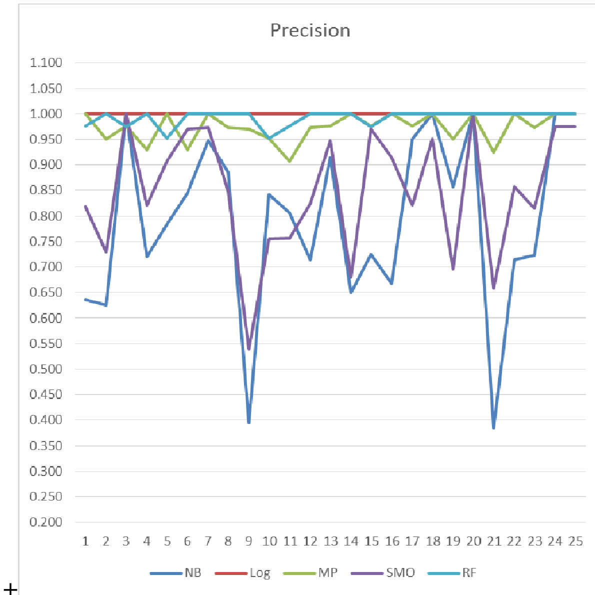

FIGURE 3.2: Graphical representation of Precision values.

Class	NB	Log	MP	SMO	RF
1	0.525	1.000	0.950	0.68	1
2	0.750	1.000	0.975	0.68	1
3	0.850	1.000	1.000	0.88	0.975
4	0.775	0.975	0.975	0.8	1
5	0.900	1.000	1.000	0.98	1
6	0.825	1.000	0.975	0.8	1
7	0.900	1.000	1.000	0.95	1
8	0.775	1.000	0.925	0.7	1
9	0.850	0.975	0.825	0.68	1
10	0.800	1.000	1.000	0.93	1
11	0.725	1.000	0.975	0.78	1
12	0.750	1.000	0.950	0.83	1
13	0.800	1.000	1.000	0.9	1
14	0.650	0.975	0.975	0.85	1
15	0.725	0.975	1.000	0.85	0.975
16	0.850	1.000	0.975	0.8	0.975
17	0.975	0.975	1.000	0.8	1
18	0.900	0.975	1.000	0.98	1
19	0.600	0.975	0.975	0.8	1
20	1.000	1.000	1.000	1	1
21	0.250	1.000	0.925	0.73	0.975

22	0.750	1.000	0.975	0.9	1
23	0.525	1.000	0.950	0.78	0.9
24	1.000	1.000	1.000	0.98	1
25	0.850	1.000	1.000	0.98	1

FIGURE 4.1: Values of Recall of each class for different classification methods.

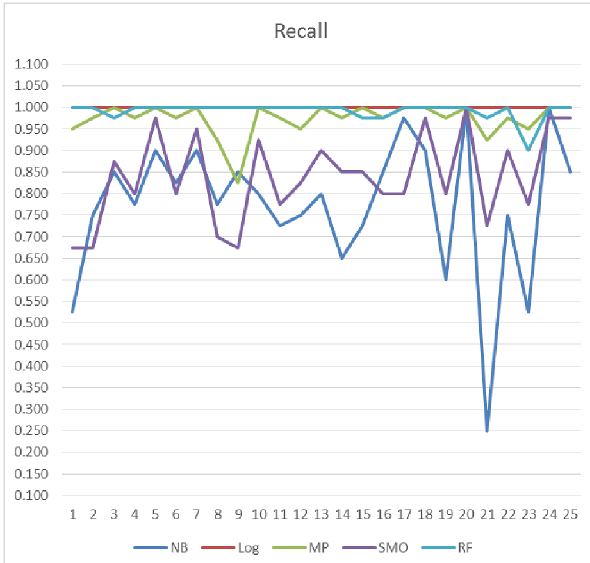

FIGURE 4.2: Graphical Representation of values of Recall.

Class	NB	Log	MP	SMO	RF
1	0.575	1.000	0.974	0.74	0.988
2	0.682	1.000	0.963	0.7	1
3	0.919	1.000	0.988	0.93	0.975
4	0.747	0.976	0.951	0.81	1
5	0.837	1.000	1.000	0.94	0.976
6	0.835	1.000	0.951	0.88	1
7	0.923	1.000	1.000	0.96	1
8	0.827	1.000	0.949	0.77	1
9	0.540	0.979	0.892	0.6	1
10	0.821	1.000	0.976	0.83	0.976
11	0.763	1.000	0.940	0.77	0.988
12	0.732	1.000	0.962	0.83	1
13	0.853	1.000	0.988	0.92	1
14	0.650	0.982	0.987	0.76	1
15	0.725	0.986	0.988	0.91	0.975
16	0.747	1.000	0.987	0.85	0.987
17	0.963	0.976	0.988	0.81	1
18	0.947	0.993	1.000	0.96	1

19	0.706	0.981	0.963	0.74	1
20	1.000	1.000	1.000	1	1
21	0.303	1.000	0.925	0.69	0.987
22	0.732	1.000	0.987	0.88	1
23	0.609	1.000	0.962	0.8	0.947
24	1.000	1.000	1.000	0.98	1
25	0.919	1.000	1.000	0.98	1

FIGURE 5.1: Values of F-measure of each class for different classification methods.

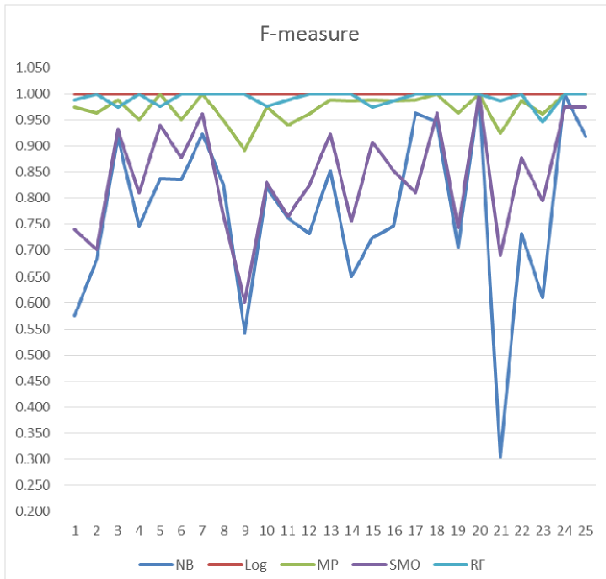

FIGURE 5.2: Graphical representation of F-measure values.

Class	**NB**	**Log**	**MP**	**SMO**	**RF**
1	0.970	1.000	0.974	0.97	1
2	0.979	1.000	0.996	0.98	1
3	0.997	1.000	1.000	1	1
4	0.984	1.000	0.999	0.99	1
5	0.995	1.000	1.000	1	1
6	0.982	1.000	0.996	0.98	1
7	0.999	1.000	1.000	1	1
8	0.991	1.000	0.985	0.98	1
9	0.977	1.000	0.964	0.97	1
10	0.995	1.000	1.000	0.99	1
11	0.983	1.000	0.999	0.98	1
12	0.986	1.000	0.995	0.99	1
13	0.996	1.000	1.000	1	1
14	0.985	1.000	0.999	0.98	1
15	0.984	1.000	1.000	0.99	1

16	0.990	1.000	0.997	0.98	1
17	0.998	1.000	1.000	0.99	1
18	1.000	1.000	1.000	1	1
19	0.978	1.000	0.998	0.98	1
20	1.000	1.000	1.000	1	1
21	0.930	1.000	0.993	0.98	1
22	0.974	1.000	0.997	0.99	1
23	0.964	1.000	0.978	0.98	1
24	1.000	1.000	1.000	1	1
25	1.000	1.000	1.000	1	1

FIGURE 6.1: Values of ROC Area of each class for different classification methods.

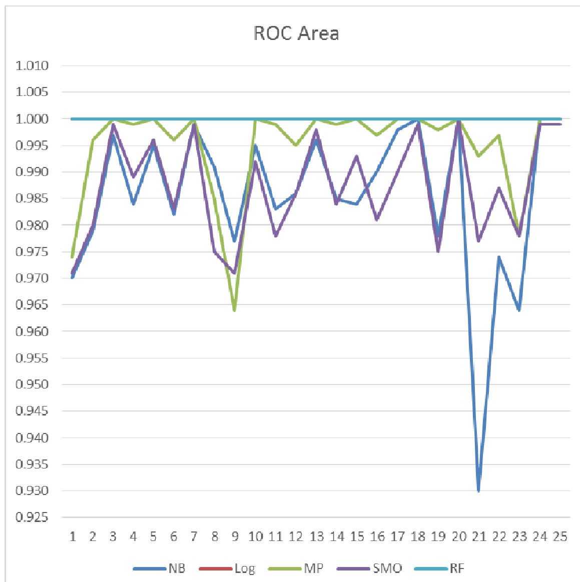

FIGURE 6.2: Graphical representation of ROC Area Values.

The following tables and diagrams pertain to the second set of experiments, i.e. with a cross validation factor of 10 in each case.

Class	MP CV10	NB CV10	Log CV 10	RF CV10	SMO CV10
1	0.725	0.5	0.725	0.675	0.525
2	0.825	0.675	0.775	0.675	0.625
3	0.925	0.775	0.975	0.875	0.825
4	0.825	0.775	0.875	0.7	0.725
5	0.900	0.9	0.925	0.8	0.9
6	0.750	0.825	0.8	0.725	0.775
7	0.950	0.85	0.975	0.925	0.925
8	0.825	0.725	0.9	0.75	0.575
9	0.625	0.85	0.725	0.65	0.675

10	0.925	0.775	0.925	0.9	0.825
11	0.875	0.7	0.9	0.85	0.75
12	0.825	0.7	0.825	0.75	0.775
13	0.975	0.775	0.975	0.825	0.85
14	0.800	0.575	0.85	0.7	0.8
15	0.925	0.675	0.9	0.825	0.85
16	0.825	0.825	0.85	0.7	0.75
17	0.850	0.95	0.975	0.925	0.7
18	0.975	0.9	0.975	0.975	0.95
19	0.875	0.575	0.9	0.675	0.725
20	1.000	1	1	1	1
21	0.675	0.225	0.75	0.45	0.575
22	0.850	0.75	0.9	0.775	0.825
23	0.775	0.425	0.8	0.625	0.7
24	0.975	1	1	0.95	0.975
25	0.925	0.825	1	0.925	0.975

FIGURE 7.1: Values of TP Rate of each class for different classification methods with cross validation 10.

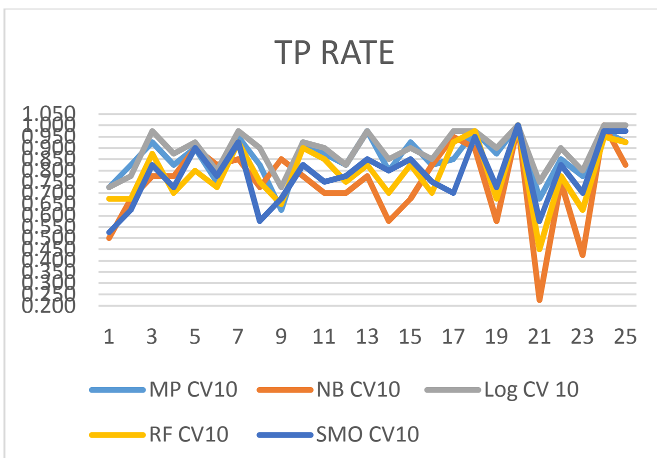

FIGURE 7.2: Graphical representation of TP Rate values with Cross Validation.

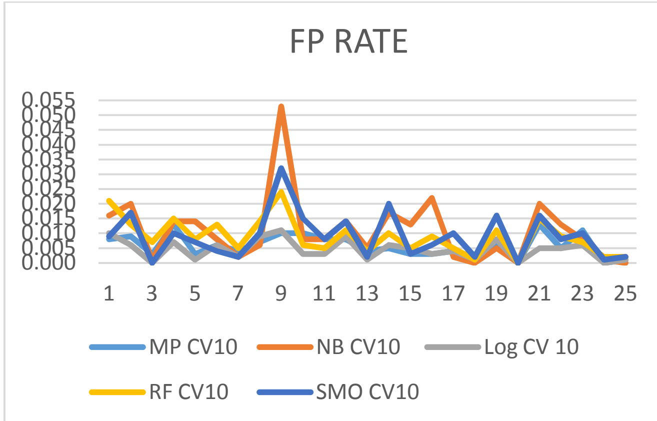

FIGURE 8.2: Graphical Representation of FP Rate values with cross validation 10.

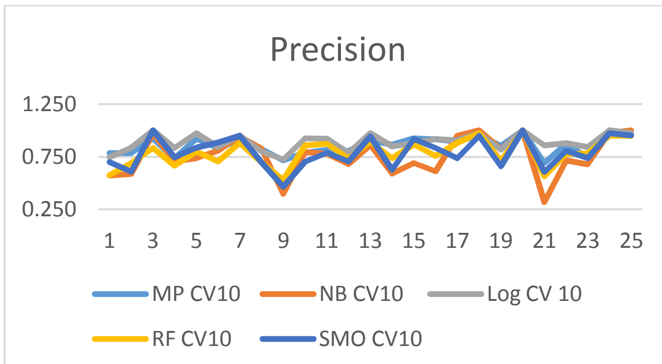

FIGURE 9.2: Graphical Representation of Precision values with cross validation 10.

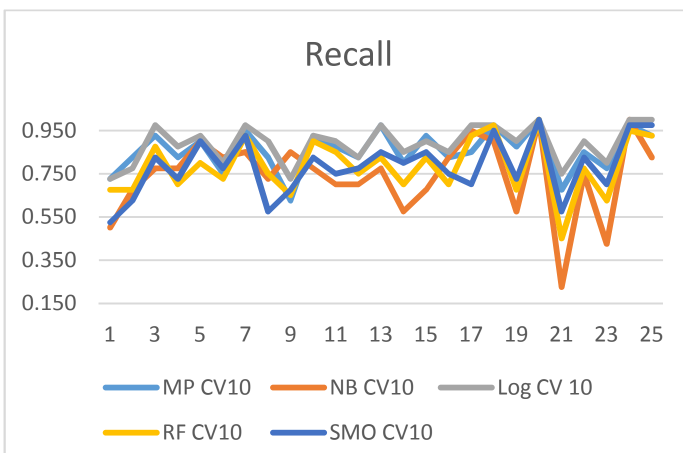

FIGURE 10.2: Graphical Representation of Recall values with cross validation 10.

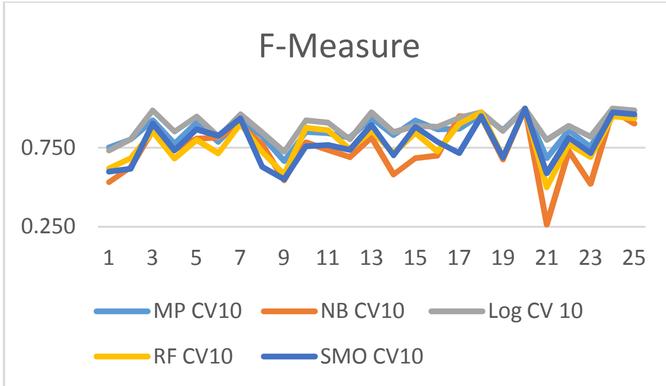

FIGURE 11.2: Graphical Representation of F-measure values with cross validation 10.

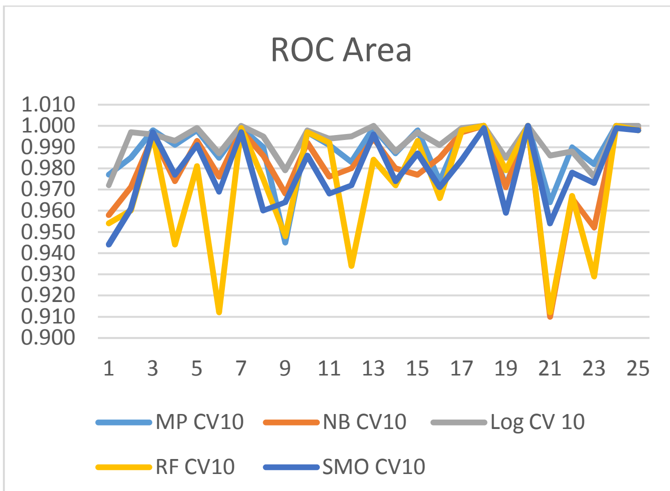

FIGURE 12.2: Graphical Representation of ROC Area values with cross validation 10.

The overall accuracy of each algorithm, considering all classes is depicted below.

Log	MP	NB	RF	SMO
99.7	97.3	77.2	99.2	83.9

FIGURE 13.1: Overall accuracy values of all classes.

5. CONCLUSION

Our comparative study provides a comprehensive analysis to Haralick features and its use in the well-known classification models. From the analysis, we can see Logistic classifier performs extremely well under all parameters, which is reflected in the combined accuracy values. It has a 99.7 percent accuracy for the trained parameters across all the classes. Random Forest Classifier performs second with respect to the rest of the classifiers. It successfully predicted all the values for most of the classes. Native Bayes performs the worst, especially with certain classes, which brings down the total accuracy achieved.

On applying cross validation with a factor of 10, we see that the accuracy decreases across all the classifiers. The different classifiers perform similarly with respect to each other as they did without cross validation. However, it can be seen that MultiLayer Perception Classifier performs slightly better than Random Forest Classifier in this case.

Apart from Native Bayes, all other methods had an accuracy of over 80 percent. Logistical and Random Forest scored above 99 percent in its accuracy. This demonstrates the power of Haralick features and its efficiency in image classification using standard classification models.

6. REFERENCES

[1] Timo Ojala, Matti Pietikainen and David Harwood, A comparative study of texture measures with classification based on feature distributions.

[2] M. Pietikainen, T. Ojala, Z. Xu; Rotation-invariant texture classification using feature distributions

[3] Eizan Miyamotol and Thomas Merryman Jr; FAST CALCULATION OF HARALICK TEXTURE FEATURES.

[4] Frank, J. (1990) Quart. Rev. Biophys. 23, 281-329.

[5] Baharak Goli and Geetha Govindan ; WEKA – A powerful free software for implementing Bio-inspired Algorithms;State Inter University Centre of Excellence in Bioinformatics, University of Kerala).

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[7] Tom M. Mitchell; Machine Learning; McGraw Hill,2010.

[8] David A. Freedsma; Statistical Models: Theory and Practice; Cambridge University Press, 2009, p. 128.

[9] Shankar Pal, Shushmita Mitra; Multilayer Perceptron, Fuzzy Sets and Classification; IEEE Transactions on Neural Networks, Vol 3, September 1992.

[10] Leo Breiman ; "Random Forests". Machine Learning 45 (1); 2001.

[11] John Platt; Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines; 1998.