International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 20 (2017) pp. 9535-9541

Inverse Fisher Discriminant Ratio based Training Set Selection for Optimal Classification Accuracy

Ashutosh Marathe[1*], Prof. (Dr.) Vibha Vyas[2] and Prof. (Dr.) Priti Rege[3]

[1,2,3] Department of Electronics & Telecommunication Engineering,
College of Engineering, 5, Wellesely Road, Shivajinagar, Pune 411005, Maharashtra, India.
**Corresponding Author*

[1]Orcid: 0000-0001-6716-602X

Abstract

Traditionally, the Fisher Discriminant Ratio (FDR) has been used to check the inter-cluster and intra-cluster distance in data points. The larger the FDR, the better the separability of data, enabling a relatively simpler classifier to perform the task with significant accuracy. In this paper, a novel Inverse Fisher Discriminant Ratio (IFDR)-based approach was used to find the optimal performance of a classifier or classifier ensemble for classification of Igneous rock into Volcanic and Plutonic Rock subfamilies. K-Nearest Neighborhood, Radial Basis Function SVM and AdaBoost with SVM as a weak classifier were used for the classification.

Random training datasets were generated out of the variety of Volcanic and Plutonic rock microstructure images. To comply with the broad need to have a training dataset that is uniformly distributed in the feature set, IFDR was calculated for each such set. Consistent with the larger IFDR, for training datasets with diverse representation from feature space, maximum accuracy was reported. This was consistent observation, in case of all classifiers and classifier ensemble. Thus, the IFDR-based approach also provided an estimate of maximum classification accuracy that can be obtained using Classifiers and classifier Ensemble. Better performance is reported for AdaBoost Classifier.

The causal relation reported between IFDR for testing the diversity of Training set and optimal accuracy for any classifier, can be used for variety of databases.

Keywords: AdaBoost, Classifier Ensemble, Haralick Features, Inverse Fisher Discriminant Ratio, Igneous Rocks, K-nn Classifier, Laws Masks, Radial Basis Function, Support Vector Machine

INTRODUCTION

Pattern classification principles have been applied in diverse areas, as the field has matured over the last two decades. Various approaches for classification of images have been proposed from time to time. Many pattern recognition approaches have been used successfully in interdisciplinary areas, such as classification of medical images and metallurgical images to name a few.

This paper focuses on a similar interdisciplinary application, i.e., classification of Igneous rock microstructure images into 2 major subcategories, namely, Volcanic rock images and Plutonic rock images. Research was carried out on handheld rock specimens, where images were captured using a camera under ambient conditions and further exploration was conducted using various supervised and unsupervised approaches [1] [2] [3] [4]. In this study, the focus is on classification of locally relevant Igneous Rock Microstructure images using a robust classifier. For this purpose, appropriate textural features are selected and classification into the appropriate subclass is carried out using a suitable classifier such as Radial Basis Function (RBF) SVM. An optimal combination is selected amongst the large number of training set combinations using an innovative Inverse Fisher Discriminant Ratio (IFDR) to calculate maximum classification accuracy. Although average classifier performance can be estimated based on techniques such as K-fold validation, [5], ratios such as IFDR help to determine the best performance in terms of % Accuracy, which can be extracted from the given classifier. Various approaches such as classifier combination using Bagging, Adaptive Boosting, Decision Trees and Random Forests [6] are applied to enhance classification accuracy.

IMAGE DATABASE AND CLASSIFICATION APPROACH

Igneous Rock and Subfamilies

The Image database is a collection of microscopic images of locally relevant Igneous rock subfamilies. The names of these types of Igneous subfamily members are shown in Table No. 1.

Table No. 1: Names of Rock Subfamilies

Sr. No.	Name of Rock Subfamily	Local Presence
1	Basalt	Predominantly Present in Western India [7]
2	Andesite	Present in Bhilwada, Chittod in Rajasthan and parts of Gujrat [8]
3	Spherulite	Present in Western Saurashtra [9]
4	Pseudotachylites	Present in Aravalli mountains and Kumaun [10]
5	Pegmatite	Present in Tumkur, Raichur, Goa and Southern Maharashtra [11]
6	Dolerite	Present in Sinnar and Sangamner Boundary, in Maharashtra [12]
7	Rhyolite	Present in Western Rajasthan [13]

The Latin word Igneous means 'Fire'. Igneous rocks originate from magma, which is formed in the deeper layers of the earth, where extremely high temperatures cause the rock to melt. This molten rock, i.e., magma is less dense than its surroundings, allowing it to rise toward the earth's surface. When the magma finds a passage to the earth's surface via a volcano in the form of lava, Volcanic Igneous rocks are formed. Due to the large difference in the surrounding atmosphere, the lava cools rapidly. This rapid crystallization produces very small grains. When the magma does not find an outlet, being in a liquid state, it seeps into crevices underneath the earth's crust, is subject to high pressure and cools very slowly, thereby forming larger grain crystals. Thus two grain size categories are formed. [14]. The classification is carried out largely by Geologists based on their expertise, taking regional Geology into consideration.

128 Igneous rock microstructure images are carefully selected, of which 64 images are of Volcanic Rocks and the remaining 64 belong to the Plutonic rock category.

Texture-based classification and Feature Selection

Grain size is chosen as the basis for classification of Igneous rock microstructures

Feature selection can be defined as the process of selecting a subset of relevant variables for use in model construction. Care is to be taken to retain the class discrimination characteristics.

For image texture analysis in [15], various textural features were proposed for image analysis. Gray level Co-occurrence Matrix (GLCM) is used as a statistical tool for establishing a relationship between inter-pixel distance and angular space. Connors et al show in [16] that Correlation, Energy, Entropy, and Homogeneity are predominantly used. For the chosen Igneous rock images, these 4 along with Contrast are used. They are defined as follows:

$$\text{Contrast} = \sum_{i,j} |i-j| P(i,j) \quad (1)$$

$$\text{Correlation} = \frac{\sum_{i=1}^{Nq} \sum_{j=1}^{Nq} P(i,j)P(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (2)$$

$$\text{Energy} = \sum_{i,j} P(i,j))^2 \quad (3)$$

$$\text{Entropy} = \sum_{i,j} P(i,j) \log P(i,j) \quad (4)$$

$$\text{Homogeneity} = \sum_{i,j} \frac{P(i,j)}{1+|i-j|} \quad (5)$$

Texture energy approach [17] is another robust texture analysis tool and, accordingly, two more features, namely, Laws Absolute Mean (Laws AM) and Laws Standard Deviation (Laws SD), were added.

Thus, the 7 features selected are (i) Contrast (ii) Correlation (iii) Energy (iv) Entropy (v) Homogeneity (vi) Laws Absolute Mean (vii) Laws Standard Deviation

By feature selection, redundant or meaningless features are discarded, so that higher generalization performance and faster classification can be realized than by the initial set of features.

The Forward selection method has been adopted. The process of forward selection begins from an empty set of features with the addition of a feature at a time. [18]

As a standard training practice, 64 images were selected using the MATLAB function randperm (n). Amongst these, half of each were from the Volcanic and Plutonic class.

To address the issue of identifying the optimal feature set, the forward Selection approach is used.

The Average False Rejection Ratio (AFRR) was calculated based on 10000 trials on various duplets. Figure No. 1 shows that the Contrast (Feature No. 1) and Laws Absolute Mean (Feature No. 6) pair produced an AFRR of 0.125. For this pair, the best and worst FRR were 0.1163 and 0.1656, respectively.

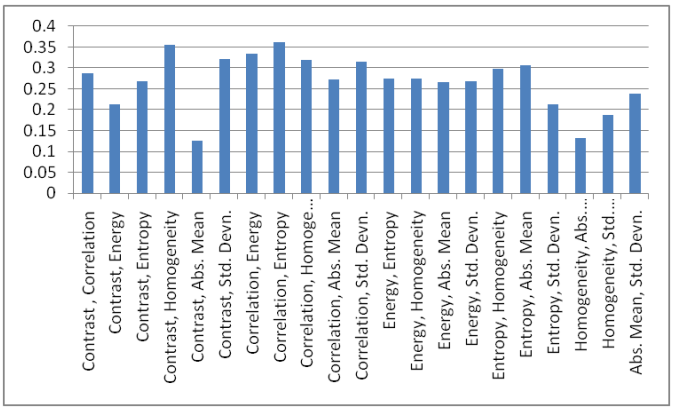

Figure 1: Average False Rejection Ratio (AFRR) for 21 duplets

Continuing with the forward selection algorithm procedure, triplets with (1,6) pair as a base were formed. For the triplet -

Laws absolute mean, Contrast and Energy, the best FRR of 0.1788 was obtained.

Further, with the use of quadruplets and so on, it was observed that the AFRR deteriorated, as shown in Fig. 2.

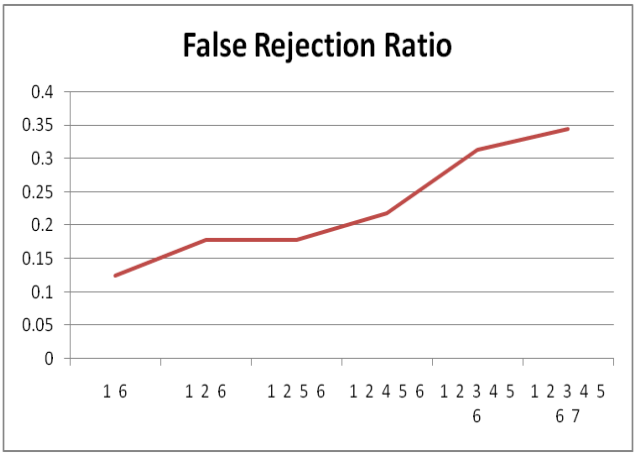

Figure 2: False Rejection Ratio Value variation

Accordingly, the duplet was considered for further work

Choice of Support Vector Machine Classifier

Support Vector machines first define basis functions that are centered on the training data points and then select a subset of it during training. One advantage of SVM is that although the training involves non-linear optimization, the objective function is convex and so the solution of the optimization problem is relatively straightforward. [19]

In general, the only available data is a finite sample of the universe U; the value of error Err (h) for a hypothesis h cannot be computed exactly. This error is first approximated by summation over the objects of the learning set, called the 'empirical risk'. [20]

Let H be the hypothesis space having Vapnik – Chevonenkis (VC) dimension d. For any probability distribution D on X {-1, 1}, with probability 1- ∂ over l random example S, any hypothesis h Є H that makes k errors on the training set has error no more than

$$\text{err (h)} \leq \frac{2k}{l} + \frac{4}{l}\left(d \log \frac{2l}{d} + \log \frac{4}{\partial}\right) \quad (6)$$

provided d ≤ l. Thus, a learning algorithm should seek to minimize the number of training errors since everything else in the bound has been fixed by the choice of the hypothesis class. This principle is known as empirical risk minimization. [21]

The ability of a hypothesis to correctly classify the data not in the training set is known as generalization.

The linear kernel SVM is discarded for the following reasons: 1. It tends to perform well if the number of features is large compared to the size of the data. In the reported work, only 2 features are chosen. 2. When linear SVM was applied, it resulted in significantly lower accuracy, although it has high generalization ability.

To obtain the best classification rate, we consider the confidence interval and empirical risk. The complex system has a high confidence interval and low empirical risk. The non-linear SVMs can reduce the empirical risk using complex systems [22]

Amongst the possible choices for non-linear SVMs, the polynomial kernel was discarded owing to its sensitivity issues.

The Sigmoid kernel is more popular in neural network-based applications. It is expressed as

$$K(X_i, X_j) = \tanh(\beta_0 X_i^T X_j + \beta_0) \quad (7)$$

Being a positive semi-definite function [23] was a constraint against choosing it.

The Radial Basis Function is typically recommended when the number of features is small. In this work, the only feature pair is chosen.

$$K(\mathbf{X_i}, \mathbf{X_j}) = \exp\left(\frac{||\mathbf{Xi}-\mathbf{Xj}||^2}{2\sigma^2}\right) \quad (8)$$

This is the default kernel chosen by SVMLib application, and traditionally it is recommended as a kernel of choice for commencing classification with non-linear SVM kernels [24].

Selection of optimal Training set

As shown in section 2.2, two features – namely Contrast and Laws Absolute Mean – have been chosen as the preferred feature pair to strike a balance between generalization of performance as well as minimization of redundancy in feature selection.

Radial Basis Function (RBF) Support Vector Machine has been selected as the preferred classifier.

Amongst the various training approaches such as Resubstitution, Hold out or Rotation method [25], the Hold out method is selected for training based on the database size and features used in the 2-class problem.

The objectives of the classification activity are as follows:

1. To estimate the 'average' performance of the classifier over the choice of randomly selected Training Dataset

2. To estimate the 'Best' performance of the classifier

For obtaining diversity in the training database, a MATLAB function randperm(n) is chosen. The function selects a random training database. 10000 such combinations are considered.

Discriminant Ratio Calculations

Traditionally, for a 1-dimensional, 2-class case, a ratio called Fisher Discriminant Ratio (FDR) is calculated as

$$\text{FDR} = (\mu_1 - \mu_2)^2 \ / \sigma_1^2 + \sigma_2^2 \quad (9)$$

where μ1 andμ2are the mean value of data points in Class 1 and Class 2, respectively, and σ1and σ2 are the variances in the 2 Classes, respectively.

FDR is typically used to quantify the separability capabilities in a primitive fashion independent of the underlying statistical distributions. [26]. The larger value of FDR indicates large between-class distances and small within-class variance. Such feature combinations are preferred since it suggests the potential use of a simple classifier, as the data constellation is such that the two classes are separated from each other, although the data points within a class are in a compact cluster.

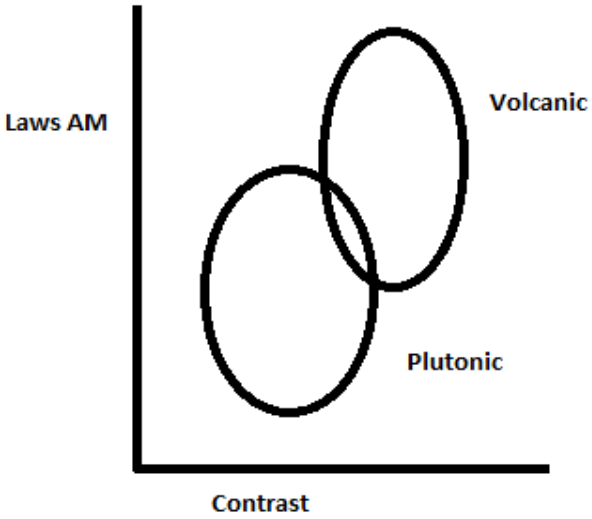

Figure 3: Placement of subfamily datasets

As far as the attributes of a good training set are concerned, it should have a uniform representation of the training sample from all the regions of the chosen sample space. [27]

In the case where the Igneous rock microstructure image classification is concerned, based on the Haralick feature – Contrast and Laws Absolute Mean (AM) feature, it is seen that owing to large grain size, the Plutonic rock images tend to occupy the bottom left area, while due to smaller grain sizes, the Volcanic rocks have a larger Contrast value and are typically present in the top right position, as shown in Figure No. 3. However, there are 'outliers,' i.e., data points belonging to one class that exhibit characteristics of the other class. Also in the case of Igneous rocks, the physical parameters that govern the process of crystallization, namely, pressure and temperature, are such that there is no 'abruptness' in a changeover from one class to another, but there is an overlapping portion. The outliers and the Overlap typically pose challenges in classification.

An Inverse Fisher Discriminant Ratio (IFDR) is calculated as shown below.

$$IFDR = (\sigma_1^2 + \sigma_2^2) / (\mu_1 - \mu_2)^2 \quad (10)$$

μ and σ have the same interpretation as mentioned in Equation No. (9)

A larger IFDR would correspond to small between-class distances and larger within-class variance. In a way, it announces a proximity and mixing of data points belonging to each class as well as a liberal spread of data points from both classes, occupying a dominant region of representing space.

CALCULATION OF IFDR AND ACCURACY

10000 random dataset combinations are prepared using MATLAB® to ensure diversity in the training datasets. Each training data set selected 32 out of 64 Volcanic Rock images and 32 out of 64 Plutonic Rock images.

For each case, a calculation of Contrast and Laws AM features was performed. The results are reported for

a) K-Nearest Neighbor approach for k= 5 as a base classifier

b) Radial Basis Function Support Vector Machine (RBFSVM) Classifier

c) Using the AdaBoost Approach for the classifier combination with RBFSVM as a base classifier.

In all 3 cases, Percentage Accuracy is calculated.

Among the randomly generated Training datasets, those sets having a biased and skewed collection of data points, it was observed that the IFDR value was small. Consequently, it was seen that the Classification Accuracy was not good.

RESULTS

For each of the 10000 iterations, values of IFDR were calculated. Table no. 2, summarizes the sample results.

Table 2: IFDR and % Accuracy Calculation - Sample Results

Sr. No	dist_ means	Var_ Plu	Var_ Vol	IFDR * 10^{-3}	% Accuracy		
					K-nn	SVM	Ada
1	27.95	0.60	0.79	1.27	0.71	0.89	0.92
2	37.91	0.65	0.78	0.72	0.68	0.86	0.9
3	54.68	0.76	0.64	0.33	0.68	0.84	0.89
4	57.00	0.70	0.65	0.28	0.68	0.85	0.86
5	67.66	0.65	0.70	0.20	0.65	0.82	0.85
6	89.10	0.66	0.75	0.13	0.66	0.82	0.85
7	118.34	0.43	0.54	0.03	0.66	0.83	0.84
8	189.87	0.35	0.68	0.02	0.63	0.84	0.84
9	213.57	0.43	0.70	0.01	0.62	0.82	0.83
10	182.44	0.48	0.38	0.01	0.58	0.79	0.81

11	258.09	0.51	0.60	0.009	0.56	0.75	0.8
12	217.79	0.34	0.53	0.008	0.56	0.75	0.81
13	270.10	0.52	0.57	0.008	0.57	0.75	0.78
14	221.33	0.30	0.53	0.007	0.53	0.72	0.75
15	247.68	0.53	0.23	0.005	0.52	0.71	0.74

Table No. 2 shows only sample results. The samples chosen are such that the results highlight a trend of increase in % Accuracy with an increase in IFDR. However, there are local exceptions where such direct proportion is not seen, while the overall direct proportionality between % Accuracy and IFDR is observed, as seen in Fig. No.4.

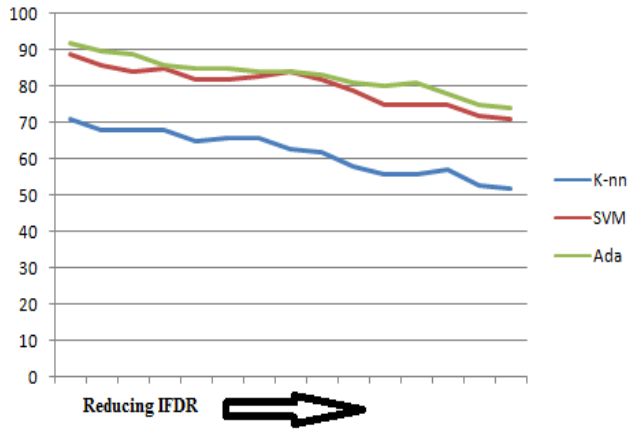

Figure 4: Relationship between IFDR and % Accuracy

Among the values shown in Table No. 2:

1. Entry at Sr. No.1 for the 'Optimal Training set', IFDR is the best and the best % Accuracy values are reported. The relevant training set is as shown in Figure No. 5.

2. Entry at Sr. No.15 for a 'highly skewed Training set', IFDR is the worst, and the least % Accuracy value is reported.

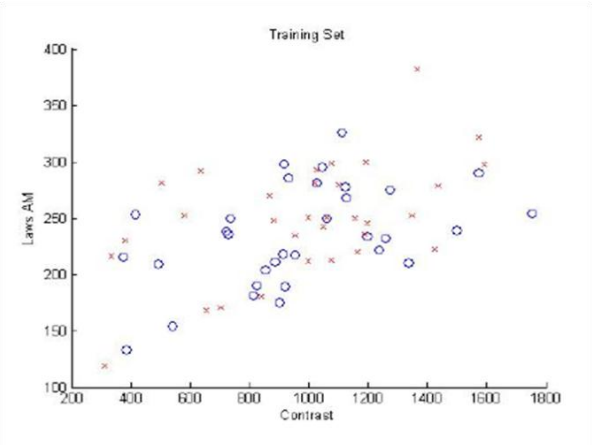

Figure 5: The Training Set associated with best IFDR

CONCLUSIONS

1. The Inverse Fisher Discriminant Ratio (IFDR) can be used as an estimation of the most optimal classifier performance.
2. The Percentage Accuracy associated with the training set giving the best IFDR is the highest value of Accuracy achievable – namely, 71% for K-nn Classifier, 89% for SVM Classifier and 92% for AdaBoost SVM Classifier.
3. The Percentage Accuracy is Maximum for AdaBoost SVM, reducing progressively for the RBF SVM and K-nn Algorithms.
4. An average Percentage Accuracy was calculated for 10000 combinations of Training set. It is 84.73 % for AdaBoost SVM, 77.96% for RBF SVM and a low average value of 58.76% for K-nn classifier in the case of Volcanic and Plutonic Rock classification.
5. Using the Classifier combination, an improvement in optimal Percentage Accuracy of around 3 % is obtained in comparison with SVM, and the Accuracy improvement is around 21% compared to K-nn Classifier. The average Percentage Accuracy improvement using Classifier combination compared with SVM Classifier is 6.77%, and that compared to K-nn Classifier is 15.97%.

The improvement in Optimal Percentage Accuracy and Average Percentage Accuracy for the Classifier Ensemble and the other two classifiers is shown in Figure No. 6.

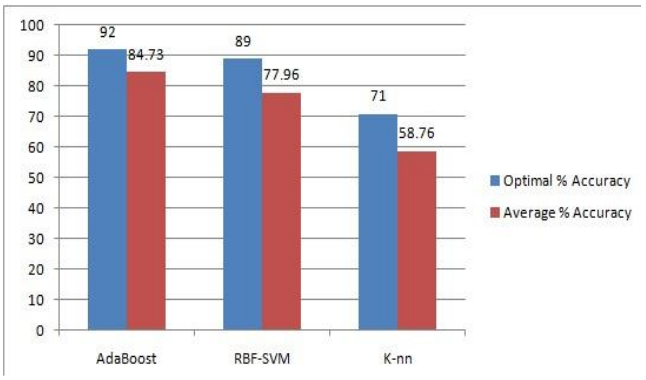

Figure 6: Optimal % Accuracy and Average % Accuracy values for Classifier Ensemble and other Classifiers

6. Not only are the Optimal Percentage Accuracy and Average % Accuracy values the best for the AdaBoost Ensemble Classifier approach, but the difference between them is the least, confirming the robustness and superiority of the Classifier ensemble in comparison with the individual classifiers. This fact is represented in Figure No. 7.

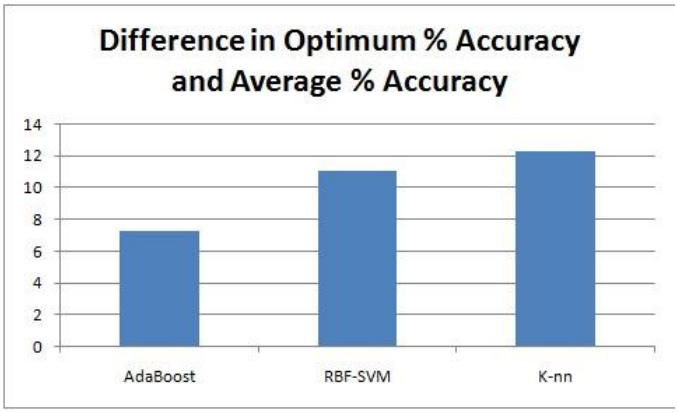

Figure 7: Difference in Optimal % Accuracy and Average % Accuracy values for Classifier Ensemble and other Classifiers

There are statistical, computational and representational reasons, substantiating an improvement in performance due to an ensemble of classifiers [28]

REFERENCES

[1] V. Shankar, J.J. Rodriquez, M.E.Gettings, 'Texture Analysis for Automated Classification of Geologic Structures' *IEEE Southwest Symposium on Image Analysis and Interpretation*. 81-85 ,2006

[2] Hang Zhou, S.T. Monterio, P.Hatherly, F.Ramos, E. Nettleton, F. Oppolzer, 'Automated rock recognition with wavelet feature space projection and Gaussian Process classification' *IEEE International Conference on Robotics and Automation (ICRA).* 4444-4450, 2010

[3] Changjing Shang , D. Barnes 'Support vector machine-based classification of rock texture images aided by efficient feature selection' *The International Joint Conference on Neural Networks (IJCNN)*. 1-8, 2012

[4] Mariusz Młynarczuk, A. Gorszczyk, B. Slipek. 'The application of pattern recognition in the automatic classification of microscopic rock images' *Computers & Geosciences*, pp.126-133, 2013

[5] E. Alpaydin '*Introduction to Machine Learning*' Second Edition, PHI Learning Pvt. Ltd. , Pg. 40, 2010

[6] B. Lantz '*Machine Learning with R*' Packt Publication, pp 337-345, 2013

[7] Geological Survey of India Report *http//www.portal.gsi.gov.in/portal/page?_pageid=127,689645*

[8] Y.H. Rao, G.S.Bharadwaj, M.Venkateshwaralu, B.R.Rao 'Magnetic and Petrographic analysis of Andesite rock bodies, Khairmalia, South of Chittourgarh, Rajasthan' *International Journal of Science, Environment and Technology (IJSET)*. Vol. 1,3. 113-124, 2013

[9] P.V.Kshirsagar, H.C.Sheth, S.J.Seaman, B.Shaikh, P. Mohite, T. Gurav, D. Chandrasekharam, 'Spherulites and thundereggs from pitchstones of the Deccan Traps: geology, petrochemistry, and emplacement environments' *Bulletin of Volcanology*. Vol. 74, 559-577, 2012

[10] K.K. Agarwal, A. Sharma, N. Jahan, C. Prakash, A. Agarwal , 'Occurrence of pseudotachylites in the vicinity of South Almora Thrust zone, Kumaun Lesser Himalaya' *Current Science*. 101,3, 431-434, 2011

[11] Geological Survey of India. Miscellaneous Publication no. 30. *Geology and the mineral resources of the states of India.* Part VII - Karnataka and Goa. 2006. http://www.portal.gsi.gov.in/gsiImages/information/misc_pub_30_karnataka_2006_wm.pdf

[12] P.D. Sabale , S,A, Meshram 'Effect of dyke structure on ground water in between Sangamner and Sinnar area: A Case study of Bhokani Dyke*' International Journal of Computational Engineering Research (IJCER)*. Vol. 2,4, 1130-1136, 2012

[13] M.S. Sisodia 'Malani rhyolite: highly eroded complex crater' *Current Science*. Vol. 101,7; 946-951, 2011

[14] D. Shelley '*Igneous and Metamorphic rocks under the Microscope'* Chapman & Hall, 1993

[15] Robert Haralick, K. Shanmugam, I. Dinstein , 'Textural features for image classification' IEEE Trans. on Systems Man Cybernetics. Vol. 3,6. 610-621, 1973

[16] Richard Conners, Charles Harlow. 'A theoretical comparison of texture algorithms' IEEE Trans. on Pattern Analysis and Machine Intelligence(PAMI). 2,3. 204-222, 1980

[17] Kenneth Laws. '*Textured image segmentation*' PhD Thesis. University of Southern California. Los Angeles. 1980

[18] Shigeo Abe, '*Support Vector Machines for Pattern Classification*' 2nd Edition Springer, pg. 333 – 339 , 2010

[19] Bishop '*Pattern Recognition and Machine Learning'* Springer , Pg. 225, 2006

[20] K.P. Soman, R. Loganathan , V. Ajay. '*Machine Learning with SVM and other Kernel Methods'* Prentice Hall India, Pg. 23 , 2009

[21] Cristianini, Shawe-Taylor '*An introduction to Support Vector Machines and other kernel based learning methods'* Cambridge University Press, Pg. 58, 2014

[22] F.Y.Shih , '*Image Processing & Pattern recognition : Fundamentals and Techniques'* Wiley, IEEE Press, Pg. 521, 2010

[23] Hsuan-Tien Lin, Chih-Jen Lin. 'A Study on Sigmoid Kernels for SVM and the Training of non-PSD Kernels by SMO-type Methods' http://home.caltech.edu/~htlin/publication/doc/tanh.pdf

[24] Rita McCue '*A Comparison of the Accuracy of Support Vector Machine and Naive Bayes Algorithms in Spam Classification*' http//classes.soe.ucsc.edu /cmps242/Fall09/proj/RitaMcCueReport.pdf

[25] L. Kuncheva '*Combining Pattern Classifiers*' John Wiley & Sons, Pg. 9 , 2009

[26] S. Theodoridis , K. Koutrumbas '*Pattern Recognition*' Academic Press , pg. 289, 2008

[27] R.O. Duda, P.E. Hart, D.G.Stork et al. '*Pattern Classification*' John Wiley & Sons pg. 5 -7, 2001

[28] T. G. Dietterich. *'Ensemble methods in machine learning'* In J. Kittler and F. Roli, editors, Multiple Classifier Systems, volume 1857 of Lecture Notes in Computer Science, Cagliari, Italy,Springer, pp. 1 –15, 2000