A Support Vector Machine and Information Gain based Classification Framework for Diabetic Retinopathy Images

M.Dharani[1,] T.Menaka[2], G.Vinodhini[3]

[1, 2, 3] (PG Scholar, CSE, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India)

ABSTRACT : *Image mining is the process of applying data analysis and discovery algorithms over large volume of image data. It has especially become popular in the fields of forensic sciences, fraud analysis and health care, for it reduces costs in time and money. It allows for the identification of natural group of patients given inputs such as symptoms and further classifies or predicts derivatives from the given data. It couples traditional manual medical data analysis with data mining methods for efficient computer assisted analysis. In this work, the concept of classifying the medical data with and without feature selection technique is discussed. The features representing the useful information about the images were extracted and fed to the mining process. The experimental results demonstrate that the SVM classifier can effectively and efficiently classify the data when compared to other classification algorithms. The information gain based attribute selection method provides the results similar to SVM classifier.*

Keywords - *Image Mining, Feature Extraction, Feature Selection, SVM.*

I. INTRODUCTION

Human medical data are difficult to mine and analyze due to the heterogeneity of medical data, socio-ethical-legal issues, statistical philosophy and special status of the medicine [1]. Since the medical data are heterogeneous data mining draws the concepts mainly from machine learning and database technology to strive towards discovering some similar structural patterns in data. Data mining is the most important technology for enabling Evidence Based Medicine, which proposes strategies to apply evidence gained from medical data for the care of individual patients [2].

Mining of bio medical image data is used to get a detailed knowledge about the specific features of the data and the way in which they are expressed in the image. It will also be used to get some non trivial conclusions and predictions on the basis of image analysis. The attributes essential for interpretation of medical data and the way in which they were used for decision making can be learnt by applying data mining methods [3].

The human medical data utilizes the same methodology of transforming raw medical data into clinically relevant information through knowledge-based data mining algorithms. This enables users to spend less time in the interaction with image volume to extract the clinical information, while supporting improved diagnostic accuracy [4].

Biomedical image processing has been an interdisciplinary research field attracting expertise from various fields such as mathematics, computer science and medicine. It has already become an important part of clinical routine. Through the new development of high technology and use of various imaging modalities, more challenges arise; for example, processing and analyzing a significant volume of images so that high quality information can be produced for disease diagnoses and treatment [5].

The recent advancements in both hardware and software coupled with higher end image processing and image vision tools, have made it possible to store huge amount of images. This makes an increase in number of images and image databases need the help of image mining techniques. Image mining is branch of data mining that is mainly focused with the process of knowledge discovery concerning digital images [6].

The rest of the paper is organized as follows. Section II describes the related works. Section III describes the proposed work. Section IV describes results and discussions. Section V describes conclusion.

II. RELATED WORKS

1. Feature Extraction using GLCM

Mohanaiah *et al.* [7] used gray level co-occurrence matrix (GLCM) to extract second order statistical texture features namely contrast, correlation, homogeneity, and entropy. Shweta Jain *et al.* [8] proposed a method for Gray Level Co-occurrence Matrix (GLCM) to extract the texture features and the co-occurrence matrices are calculated for four directions. Fritz Albregtsen *et al.* [9] computed the texture features by the statistical distribution of observed combinations of intensities at specified positions relative to each other. Alaa Eleyan *et al.* [10] analysed GLCM by converting the matrix into a vector that can be used in the classification process.

2.Classification techniques

Kun Che Lu *et al.* [11] identified decision tree induction to realize relationships between attributes and the target label, and to construct a model for a given image dataset. Angelos Tzotsos *et al.* [12] discussed Support Vector Machine for the classification of datasets of higher dimensionality and it worked the best when compared to other machine learning algorithms. Radu Timofte *et al.* [13] proposed Naive Bayes Nearest Neighbor as a powerful, approach for image classification since the vector quantization step was not used. D S Guru *et al*. [14] used k-nn so that patterns closer to each other in the feature space are likely to belong to the class of the same pattern.

3. Feature Selection

Andreas G.K. Janecek *et al*. [15] described feature subset selection in his work where a subset of the original attributes is extracted and approaches such as filters and wrappers were implemented. Susana Eyheramendy *et al*. [16] discussed a new feature selection technique based on score and the value assigned to the score of each feature interprets Bayesian networks. YongSeog Kim *et al*. [17] proposed feature selection to choose a subset of input variables by eliminating features with little or no predictive information. M. Dash *et al*. [18] anayzed feature selection to search through the subsets of features and the best one among the competing 2*N* candidate subsets was found based on some evaluation function.

4. Information Gain based Feature Selection

Nicolette Nicolosi et al. [19] stated that the information gain is a method that measures the decrease in entropy when the feature is given and it can generalize to any number of classes. Susana Eyheramendy *et al.* [16] discussed a new feature selection technique based on score and the value assigned to the score of each feature interprets Bayesian networks. Sanmay Das *et al.* [20] proposed filter and wrapper methods of feature selection and proposed a new hybrid algorithm to boost and incorporate into a fast filter method for feature selection. Shailendra Singh *et al.* [21] described the filter phase to select the features with highest information gain.

This study proposes a biomedical image mining framework enabling the image processing tasks that extracts useful information from the images. The use of image processing operators provides flexibility in handling and processing the images. The major requirement of the framework is the use of a combination of *image processing* and *image mining* functionalities in a research environment. The first objective was met by the use of MATLAB [22] while the second objective has been accomplished by the coupling of MATLAB and WEKA [23] platforms. The proposed framework is developed to solve the problem of mining the extracted features and using the information in decision making.

III. PROPOSED FRAMEWORK

1 .System description

The proposed method includes image pre-processing, feature extraction, image mining and feature selection. The overview of the proposed system is illustrated in fig. 1. The images acquired were subjected to preprocessing and feature extraction steps. The preprocessing includes grayscale conversion, image cropping, dimensionality reduction and histogram equalization. The canny edge detection technique is used to identify the edges of the objects in the images. The feature extraction technique involves Gray Level Co-occurrence matrix (GLCM) [24].These features are extracted and stored separately in a table. The feature values are classified using various algorithms with and without feature selection. The class precision and class recall measures the accuracy of these algorithms and the comparison table for these measures is created.

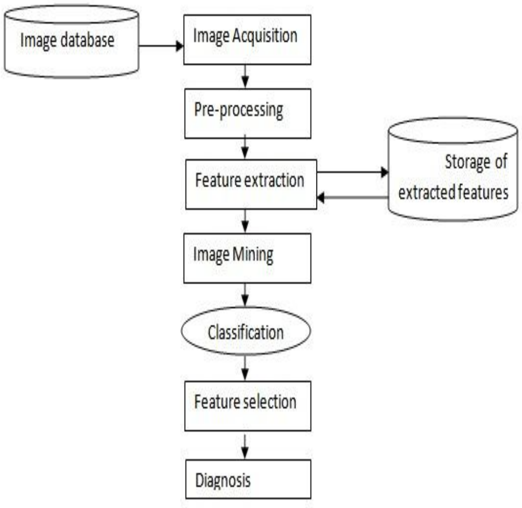


Fig. 1 Overview of the proposed system

1.1 Image Dataset

The utilized image dataset was obtained from Kuopio university hospital. The database consists of 89 colour fundus images where 84 of them contains mild non proliferative signs of diabetic retinopathy whereas 5 are normal images which do not contain any signs of diabetic retinopathy [25]. The abnormalities present in the images are relatively small since they appear near the macula they are considered to threaten the eyesight. Fig. 2 indicates the dataset image.

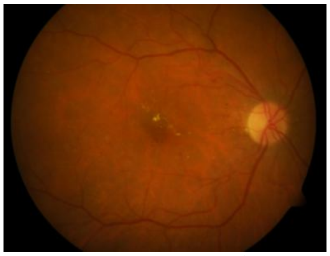


Fig. 2. DIABRETDB1 fundus image

1.2 Preprocessing

The main objective of preprocessing is to enhance the quality of the image data by improving the required image features for further processing. The preprocessing technique eliminates the irrelevant, inconsistent and incomplete data present in the images. In order to enhance the image quality and to make the feature extraction phase more accurate and reliable, pre-processing is necessary.

1.2.1 Gray Scale Conversion

The original image was true color RGB image. It is converted to grayscale image by considering the R, G, B color values of each pixel and the output is obtained by making it as a single value reflecting the brightness of that pixel. The method used in this work was simply averaging the RGB values of each pixel .

1.2.2 Histogram Equalization

When the images are digitized, noise could occur, detecting its peak value, and then adjusting each image so that its histogram peak is aligned to the average histogram. The brightness of all pixels in the image is readjusted ranged from 0 to 255.

1.2.3 Edge detection

Edge detection refers to the process of identifying the sharp image brightness changes in the points and is organized into a set of curved lines called edges. It is a fundamental tool in image processing, machine vision and computer vision. In this work the operator used for edge detection is Canny.

1.3 Feature extraction

Feature extraction is a special when the input data to be processed is too large and redundant. The transformation of input data into a reduced representation of set of features is called *feature extraction*. The main criterion in feature extraction is to carefully choose the attributes providing useful information. The feature extraction needs only a small amount of memory and less computation power while describing the data with a higher accuracy. In this work, the features extracted were mean, standard deviation, correlation, angular second moment, inverse difference moment, contrast, entropy, minimum and maximum gray scale value, mode, ROI (height, width and percentage),area, median, kurtosis,skewness, minimum and maximum value of histogram, area fraction, centroid , angle and centre of mass [26].

1.4 Image Mining

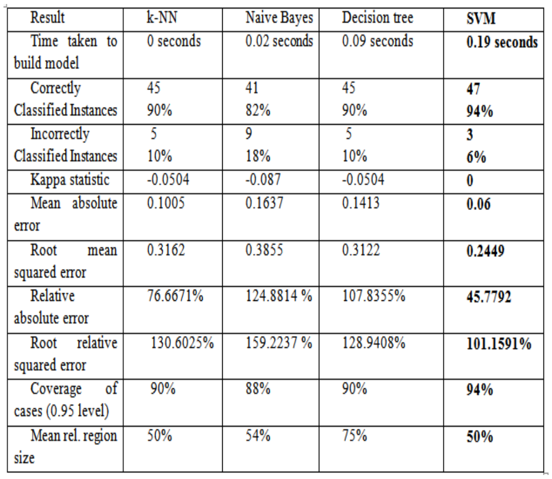
Image mining is the process of discovering valuable information from a large amount of data. The features that are extracted are stored in a table. The features were fed to the classification algorithms and were mined. The result obtained shows the class precision and recall for the algorithms. Some of the algorithms used were SVM, Decision tree, K-Nearest Neighbor and Naives Bayes [27], [28], [29], [30].

1.5 Feature selection

Feature selection is the process of choosing a subset of input variables by eliminating those with little or no useful information. It adopts both supervised and unsupervised learning.

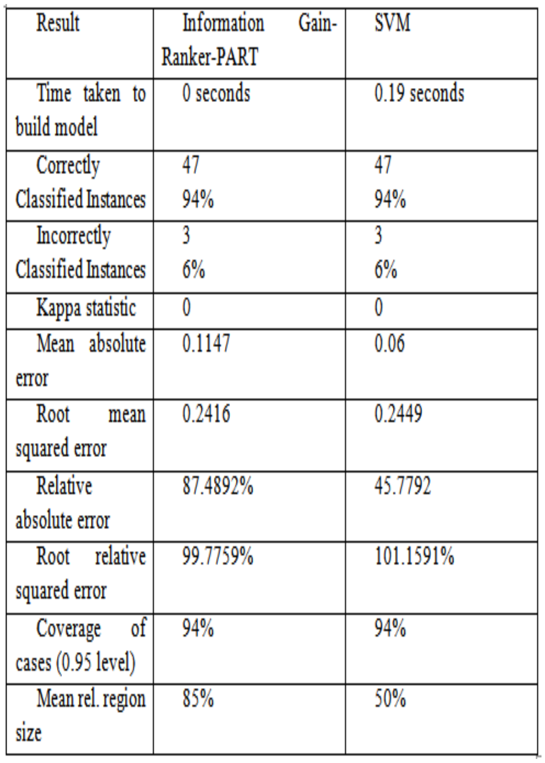
IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, an experimental analysis was conducted to evaluate the performance of proposed data mining algorithms. The performance was evaluated on diabetic retinopathy datasets by comparing the measures for the classification algorithms. The features were selected by attribute construction method using information gain and were made to classify. Finally the accuracy of the proposed work for mining the dataset is examined for classification with and without feature selection techniques.



Result	k-NN	Naive Bayes	Decision tree	SVM
Time taken to build model	0 seconds	0.02 seconds	0.09 seconds	**0.19 seconds**
Correctly Classified Instances	45 90%	41 82%	45 90%	**47** **94%**
Incorrectly Classified Instances	5 10%	9 18%	5 10%	**3** **6%**
Kappa statistic	-0.0504	-0.087	-0.0504	**0**
Mean absolute error	0.1005	0.1637	0.1413	**0.06**
Root mean squared error	0.3162	0.3855	0.3122	**0.2449**
Relative absolute error	76.6671%	124.8814 %	107.8355%	**45.7792**
Root relative squared error	130.6025%	159.2237 %	128.9408%	**101.1591%**
Coverage of cases (0.95 level)	90%	88%	90%	**94%**
Mean rel. region size	50%	54%	75%	**50%**

Table. 1 Comparison results of efficiency of the classification algorithms



Result	Information Gain-Ranker-PART	SVM
Time taken to build model	0 seconds	0.19 seconds
Correctly Classified Instances	47 94%	47 94%
Incorrectly Classified Instances	3 6%	3 6%
Kappa statistic	0	0
Mean absolute error	0.1147	0.06
Root mean squared error	0.2416	0.2449
Relative absolute error	87.4892%	45.7792
Root relative squared error	99.7759%	101.1591%
Coverage of cases (0.95 level)	94%	94%
Mean rel. region size	85%	50%

Table. 2 Efficiency results for image mining with and without feature selection

The results from the Table 1 indicate that the K-NN classifier takes less time to build the model. The accuracy of correctly and incorrectly classifying the instances was higher in SVM classifier when compared to other classification algorithms. The problem of predicting the true class and the error values is higher in SVM classifier.

When comparing efficiency results of SVM classifier and Information gain based feature selection in Table 2, SVM classifier provides less mean absolute error and root mean squared error values.

Fig. 3. Indicates the cost/benefit analysis result for support vector machine classification algorithm. The result demonstrates that the cost was decreased when compared to other algorithms.

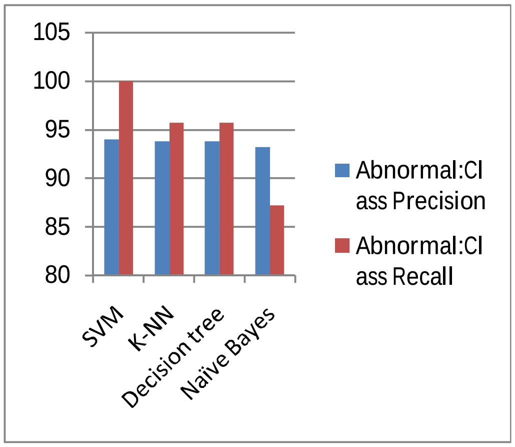


Fig. 3. Cost/Benefit Analysis for SVM

V.CONCLUSION

The image mining framework collaborated with image processing operators has been developed and the performance is evaluated. The grayscale images were histogram equalized and the edges were detected. The features were extracted from the edges and were provided as input to data mining algorithms. The proposed work compares the efficiency and accuracy of the classification algorithms. The results were obtained for classification algorithms with and without feature selection. The results can assist as a second option for physicians in diagnosis. In future work, the attributes will be selected based on the varying evaluator methods and searching techniques.

REFERENCES

Journal Papers:

[1] Krzysztof J. Cios, G. William Moore,"Uniqueness of medical data mining", *Artificial Intelligence in Medicine 26 (2002) 1–24.*

[6] A.Hema1, E.Annasaro2," A survey in need of image mining techniques", *International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 2, February 2013 ISSN (Print): 2319-5940.*

[7] P. Mohanaiah*, P. Sathyanarayana**, L. GuruKumar***," Image Texture Feature Extraction Using GLCMApproach",*International Journal of Scientific and Research Publications, Volume 3, Issue 5, May 2013 1 ISSN 2250-3153.*

[8] Shweta Jain," Brain Cancer Classification Using GLCM Based Feature Extraction in Artificial Neural Network" *International Journal of Computer Science & Engineering Technology (IJCSET).*

[10] Alaa ELEYAN1, Hasan DEM˙IREL2," Co-occurrence matrix and its statistical features as a new approach for face recognition" *Turk J Elec Eng & Comp Sci, Vol.19, No.1, 2011, c_T¨UB˙ITAK doi:10.3906/elk-0906-27.*

[11] Kun-Che Lu And Don-Lin Yang," Image Processing and Image Mining using Decision Trees*", *Journal of information science and engineering 25, 989-1003 (2009).*

[14] D S Guru, Y. H. Sharath, S. Manjunath," Texture Features and KNN in Classification of Flower Images", *IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition"RTIPPR, 2010.*

[21] Shailendra Singh, Sanjay Silakari," An ensemble approach for feature selection of Cyber Attack Dataset" *(IJCSIS) International Journal of Computer Science and Information Security,*
Vol. 6, No. 2, 2009.

[24] Dr. H.B.Kekre, Sudeep D. Thepade, Tanuja K. Sarode and Vashali Suryawanshi," Image Retrieval using Texture Features extracted from GLCM, LBG and KPE" *International Journal of Computer Theory and Engineering, Vol. 2, No. 5, October, 2010 1793-8201.*

[26] Theodosis Goudas, Aristotle Chatziioannou,"A collaborative biomedical image-mining framework:application on the image analysis of microscopic kidney biopsies", *IEEE Journal of Biomedical and health informatics, Vol. 17,No.1, January 2013.*

[27] M. K. Ghose , Ratika Pradhan, Sucheta Sushan Ghose," Decision Tree Classification of Remotely Sensed Satellite Data using Spectral Separability Matrix", *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 1, No.5, November 2010.*

[28] Miss. Mayanka b. Khuman," Classification of remote sensing data using k-NN method", Journal of Information, Knowledge and Research in Electronics and Communication Engineering.

[30] Olivier Chapelle, Patrick Haffner, and Vladimir N. Vapnik," Support Vector Machines for
Histogram-Based Image Classification", Ieee Transactions on Neural Networks, vol. 10, no. 5, september 1999.

Theses:

[4] R. Bharat Rao, Glenn Fung, Balaji Krishnapuram, Jinbo Bi, Murat Dundar, Vikas Raykar, Shipeng Yu, Sriram Krishnan, Xiang Zhou, Arun Krishnan, Marcos Salganicoff, Luca Bogoni, Matthias Wolf, Anna Jerebko, Jonathan Stoeckel, "Mining Medical Images" *Image and Knowledge Management-CAD and Knowledge Solutions (IKM-CKS) Siemens Medical Solutions USA, Inc., 51 Valley Stream Parkway, Malvern, PA-19355.*

[5] Hongmei Zhu, University of Calgary," Medical Image Processing Overview".

[9] Fritz Albregtsen," Statistical Texture Measures Computed from Gray Level Coocurrence Matrices" *November 5, 2008.*

[12] Angelos Tzotsos*," A support vector machine approach for object based image analysis", Commission IV, WG IV/4 – Commission VIII, WG VIII/11.

[13] Radu Timofte, Tinne Tuytelaars, and Luc Van Gool," Naive Bayes Image Classification: beyond Nearest Neighbours".

[17] YongSeog Kim, W. Nick Street, and Filippo Menczer," Feature Selection in Data Mining".

[18] M. Dash 1, H. Liu2," Feature Selection for Classification" *Intelligent Data Analysis 1 (1997) 131–156.*

[19] Nicolette Nicolosi," Feature Selection Methods for Text Classification" November 7, 2008.

[20] Sanmay Das," Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection".

[22] Abouzar Eslami and Emadodin Fatemizadeh," Insight to Matlab Image Processing Toolbox".

[23] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer
Peter Reutemann, Ian H. Witten," The WEKA Data Mining Software: An Update".

[25] Tomi Kauppi,Valentina Kalesykiene,Joni-Kristian Kamarainen,Lasse Lensu,Iiris Sorri,Asta Raninen,Raija Voutilainen,Hannu Uusitalo,Heikki Kalviainen and Juhani Pietila, "DIABRET1DB diabetic retinopathy database and evaluation protocol".

[29] YU Xin, Zheng Zhaobao, Zhang Haitao and Ye Zhiwei ," Texture classification of aerial image based on bayesian network augmented Naive Bayes".

Proceedings Papers:

[2] Payam Homayounfar , Mieczyslaw L. Owoc," Data Mining Research Trends in Computerized Patient Records", *Proceedings of the Federated Conference on Computer Science and Information Systems pp. 133–139 ISBN 978-83-60810-22-4.*

[3] Petra Perner, "Mining Knowledge in Medical Image Databases", *In Data Mining and Knowledge Discovery: Theory, Tools and Technology, Belur VDasarathy (eds.), Proceedings of SPIE Vol. 4057 (2000),359-369.*

[15] Andreas G.K. Janecek, Wilfried N. Gansterer, Michael A. Demel, Gerhard F. Ecker," On the Relationship Between Feature Selection and Classification Accuracy" *JMLR: Workshop and Conference Proceedings 4: 90-105, New challenges for feature selection.*

[16] Susana Eyheramendy, David Madigan," A novel feature selection score for text categorization" *Proceedings of the Workshop on Feature Selection for Data Mining:*
Interfacing Machine Learning and Statistics in conjunction with the 2005 SIAM International Conference on Data MiningApril 23, 2005.