Colon Cancer Detection Based on Structural and Statistical Pattern Recognition

Beema Akbar, Varun P Gopi, V Suresh Babu

Abstract—Colon cancer causes the deaths of about half a million people every year. The common method of its detection is histopathological tissue analysis, it leads to tiredness and workload to the pathologist. A novel method is proposed that combines both structural and statistical pattern recognition used for the detection of colon cancer. This paper presents a comparison among the different classifiers such as Multilayer Perception (MLP), Sequential Minimal Optimization (SMO), Bayesian Logistic Regression (BLR) and k-star by using classification accuracy and error rate based on the percentage split method. The result shows that the best algorithm in WEKA is MLP classifier with an accuracy of 83.333% and kappa statistics is 0.625. The MLP classifier which has a lower error rate, will be preferred as more powerful classification capability.

Keywords—Colon cancer, histopathological image, structural and statistical pattern recognition, Multilayer perception.

I. Introduction

CANCER is one of the most important health problems that threat the human life. Colon is one major constituent of large intestine, and its cancer is a major reason of deaths in western and industrialized world [1]. There are many reasons of colon cancer like chain smoking, increasing age such as age above 50 years, family history of colon cancer, low intake of fruits, and heavy intake of red meat and fats.

In the human body, tissues are characterized with the organization of their component. Neoplastic diseases including cancer cause changes in these organizations [2]. Thus, the correct identification of the deformations in the structures and their accurate quantification are critical for precise modelling of cancer. Traditionally, colon cancer is diagnosed using microscopic analysis of histopathological colon samples. Histological analysis is performed by examining a thin slice (section) of tissue under a light (optical) or electron microscope. After a sequence of technical procedures for tissue preparation (i.e., fixation, dehydration, clearing, infiltration, embedding, sectioning and staining) histology images can be produced by different imaging techniques based on which manual or automated analysis can be conducted to detect diseased tissues. In such an examination, pathologists observe the colon samples under microscope depending upon the level of organizational changes they observe in tissues. But, the manual examination has a few limitations [2]. This examination is time consuming and mainly relies on the visual interpretation, and thus it may lead to a considerable amount of intra and inter observer variability. Such vulnerabilities in the manual process result in need of automatic colon cancer diagnostic techniques.

Normal colon tissues have well defined organizational structure [2]. However, this arrangement varies in case of cancer. Variation usually depends upon the cancer stage. Initial cancer stages deform the cells very little and harder to detect. On the other hand, advanced stages significantly deform the cells, thereby making their detection easier. Fig.1 shows a histopathological image of a colon tissue stained with hematoxylin and eosin. Staining is employed to give both contrasts to the tissues as well as highlighting particular features of interest. Hematoxylin and Eosin (H & E) stain is the most commonly used light microscopical stain. Hematoxylin is a basic dye, which stains nuclei blue due to an affinity to nucleic acids in the cell nucleus, Eosin is an acidic dye, stains cytoplasm pink. Fig. 2 shows constituents of a normal colon tissue. It consists of three components (epithelial cells, nonepithelial cells, and lumen) of normal colon tissue. Epithelial cells usually surround lumen and form glandular structure, whereas nonepithelial cells called stroma lie in between these structures. Fig. 3 shows the malignant colon tissue. Deformation introduced by cancer is clearly visible in Fig. 3.

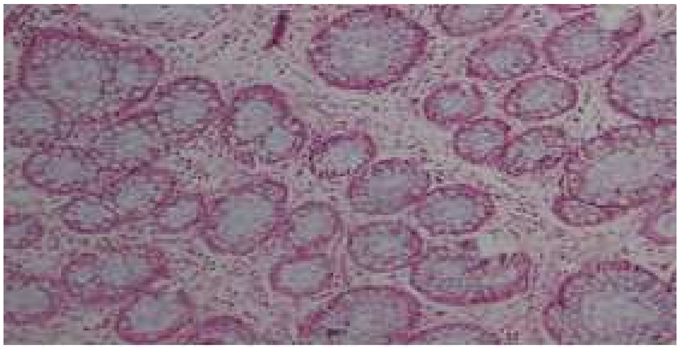

Fig. 1: A histopathological image of a colon tissue stained with hematoxylin and eosin

Remainder of this paper is organized as follows: Section II presents a detailed insight into existing colon cancer detection techniques. Section III provides a proposed method for the detection of colon cancer. Section IV & V reports the experimental results and result analysis of the classification techniques. Finally, Section VI concludes the paper.

Beema Akbar is with the Department of Electronics and Communication Engineering, Government Engineering College Wayanad, Kerala, India e-mail: nimishaelsa@yahoo.co.in

Varun P Gopi is with the Department of Electronics and Communication Engineering, Government Engineering College Wayanad, Kerala, India e-mail: varunpg@gecwyd.ac.in

V Suresh Babu is with the Department of Electronics and Communication Engineering, College of Engineering Trivandrum, Kerala, India e-mail: vsbsreeragam@gmail.com

978-1-4788-7225-8/15/$31.00 ©2015 IEEE

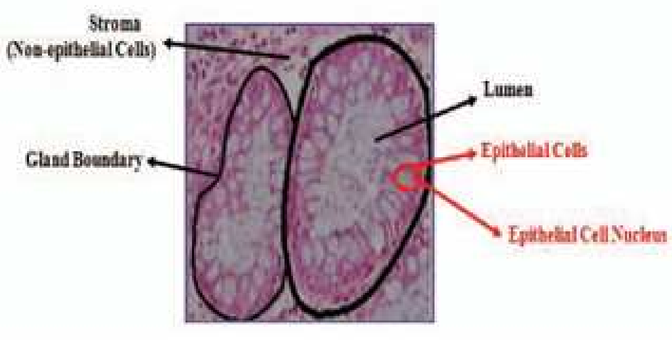

Fig. 2: Constituents of normal colon tissue

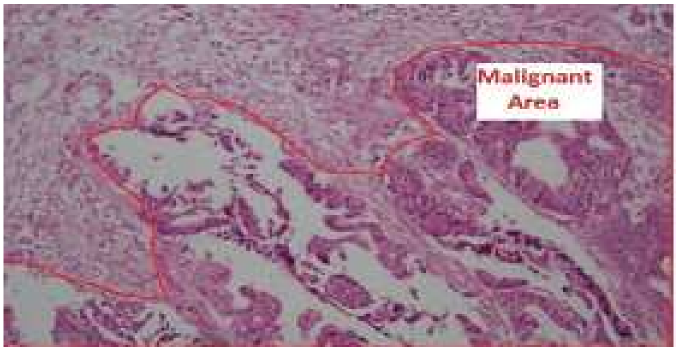

Fig. 3: Malignant colon tissue

II. Related Work

Texture, hyperspectral and Object-Oriented (OO) texture analysis based techniques work on images. A few techniques, which work on colon biopsy images but are not well established. Colon biopsy is the removal and examination of tissue, cells or fluid from the colon. It is an exam that tests tissue samples from the colon. The doctor is looking for abnormal tissues, such as cancerous or pre-cancerous cells. The only way to define, if the abnormality is cancerous is by extracting tissue and taking a look at it under a microscope. Serum analysis based techniques analyse physical sample for cancer detection therefore, it have been named physical sample based colon cancer detection techniques.

A. Texture Analysis Based Technique

Texture is a combination of repeated patterns with regular/irregular frequency. Entropy and correlation have been commonly used to quantize texture. Esgiar et al. [4] proposed a method of colon cancer detection by using textural features. In their work, original colon biopsy images are divided into four subimages. Gray Level Co occurrence Matrix (GLCM) is then calculated for each subimage. Normalized GLCM is used to determine textural features of angular second moment, contrast, correlation, inverse difference moment, dissimilarity and entropy. The reported classification accuracy is 90.2% for a combination of correlation and entropy by using Linear Discriminate Analysis (LDA) classifier. But it has a smaller data set and lengthy data acquisition method. Further, Kalkan et al. [5] combined texture and structural features to classify colon samples into normal, precancerous (adenomatous and inflamed) and malignant classes. Logistic regression classifier with equal class priorities is applied for classification. 77.29, 82.25, 76.08 and 66.86% classification accuracy is for adenomatous, malignant, inflamed, and normal classes respectively. All the techniques [4]-[5] classify colon samples into normal and malignant classes but none of them distinguishes cancer grades.

B. OO Texture Analysis Based Technique

This technique exploit background knowledge about size and spatial distribution of colon tissue components for segmentation and classification of colon biopsy images. Demir et al. [3] proposed a new gland segmentation algorithm that relies on decomposing the image into a set of primitive objects (nucleus and lumen objects) and then making use of the organizational properties of these objects instead of using the pixel based information alone. Region growing process leads to 82.57% average segmentation accuracy on the test set and that this accuracy increases to 87.59% after the false gland elimination step. This proposed object based algorithm significantly improves the segmentation performance of its pixel based counterparts. Segmentation techniques [3] only segment images, whereas classification techniques classify samples into different classes. But there is no cancer grading capability techniques. It has a smaller data set and it is computationally expensive.

C. Hyperspectral Analysis Based Technique

Hyperspectral analysis based techniques operate on selected spectral bands of colon biopsy images, and identify normal and malignant tissues. Masood et al. [6] have used GLCM and morphological features for colon tissue classification. Three distinct phases of their technique are segmentation, feature extraction, and classification. In segmentation phase dimensionality of 3D cubes of hyperspectral image data is reduced. Imaging data is then divided into four clusters of nuclei, cytoplasm, glands, and stroma by using k-means. Principal Component Analysis (PCA) and LDA are used to characterize images on the basis of morphological features and a maximum of 84% accuracy is achieved. The reported classification accuracy is 90%. But the spectral data is not easily available to histopathologists.

D. Laser Induced Fluorescence (Blood serum) Analysis Based Technique

This technique analyze the Raman spectrum of blood serum. Three well distinguished peaks are indicator of normal blood serum, whereas irregular peaks or absence of peak shows cancer. X. Li et al. [7] used laser induced fluorescence and Raman spectroscopy method for the detection of cancer. Cancer changes chemical composition of different ingredients in blood serum. Consequently, Raman spectrum of malignant serum heavily deviates from its normal counterpart. Techniques, based on laser induced fluorescence and Raman spectroscopy exploit such differences in blood serum and resultant Raman spectra for detection of colon cancer. The

978-1-4788-7225-8/15/$31.00 ©2015 IEEE

reported classification accuracy is 83.5% for a data set of 65 samples. Laser Induced Fluorescence (LIF) based techniques [7] have smaller accuracy, because equipment is quite delicate and a minute human error leads to wrong results.

III. Proposed Method

Cancer causes deviations in the distribution of cells, leading to changes in biological structures that they form. Correct localization and characterization of these structures are crucial for accurate cancer diagnosis and grading. A new hybrid model [8] is proposed which combines both structural and statistical pattern recognition techniques to locate and characterize the biological structures in a tissue image for tissue quantification. A block diagram of the proposed method is shown in Fig. 4.

A. Tissue Graph Generation

A tissue image is modelled with an attributed graph $G = \{V, E, \mu\}$ where V is a set of nodes, E is a set of edges, and μ is a mapping function that maps each node into an attribute label. This graph representation relies on locating the tissue components in the image, identifying them as the graph nodes, and assigning the graph edges between these nodes based on their spatial distribution [10]-[11]. However, as the exact localization of the components emerges a difficult segmentation problem, an approximation is used that defines circular objects to represent the components. In order to define these objects, the image pixels are quantified into two groups: nucleus pixels and non-nucleus pixels. For that, the hematoxylin stain is separated using the deconvolution method and threshold it with the Otsu method [12]. This method tries to evaluate the goodness of the threshold and automatically selects an optimal threshold according to a certain criterion. It is based on the assumption that well threshold classes would be separated in gray-levels and conversely, that a threshold, giving the best separation of several classes (assumed to correspond to uniform gray-levels) should be the best classification threshold. Then, on each group of the pixels, a set of circular objects is defined.

This approximation gives two groups of objects: one group defined in the nucleus pixels and the other defined on the non-nucleus (whiter) pixels. These groups are referred to as the nucleus and white objects. In this approximate representation, there is not always a one-to-one correspondence between the components and the objects. For instance, while a nucleus component typically corresponds to a single nucleus object, a Lumen component usually corresponds to many white objects forming a clique. Thus, in addition to the types of the objects, their spatial relations are also important for this tissue representation. Graphs can model such relations [11]. After defining the objects as the graph nodes, their spatial relation is encoded by constructing a tissue graph using Delaunay triangulation [13].

B. Query Graph Generation

Query graphs are the subgraphs that correspond to normal gland structures in an image. To define a query graph on the tissue graph of a given image, a seed node (object) is selected and expand it on the tissue graph using the breadth first search (BFS) algorithm until a particular depth is reached. Then, take the visited nodes and the edges between these nodes to generate the query graph. In this procedure, the seed node and the depth are manually selected, considering the corresponding gland structure in the image.

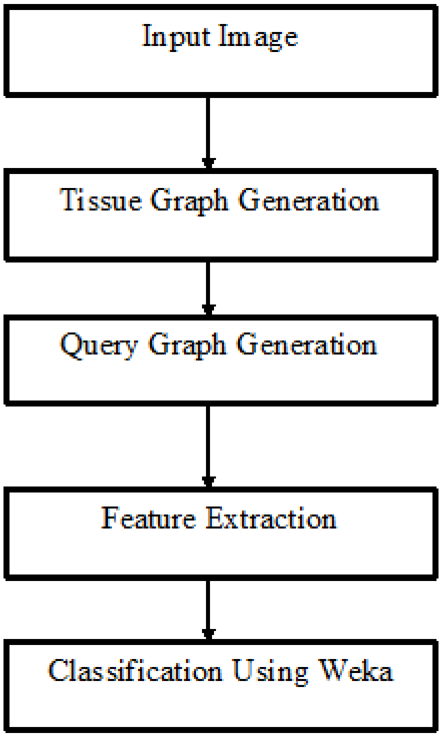

Fig. 4: Proposed Method

The localization of key regions in an image includes a search process. This process compares each query graph with subgraphs generated from the tissue graph of the image and locate the ones that are the most similar to this query graph. The regions corresponding to the located subgraphs are then considered as the key regions. Since a query graph is generated as to represent a normal gland, the located subgraphs are expected to correspond to the regions that have the highest probability of belonging to a normal gland. Typically, the subgraphs located on a normal tissue image are more similar to the query graph than those located on a cancerous tissue image. Thus, the similarity levels of the located subgraphs together with the features extracted from their corresponding key regions are used to classify the tissue image.

C. Feature Extraction

In order for the pattern recognition process to be tractable, it is necessary to represent patterns into some mathematical or analytical model. The model should convert patterns into features or measurable values, which are condensed representations of the patterns, containing only salient information [11]. Gray Level Co-Occurrence Matrix (GLCM) has proved to be a popular statistical method of extracting

978-1-4788-7225-8/15/$31.00 ©2015 IEEE

textural feature from images. According to co-occurrence matrix, Haralick defines fourteen textural features measured from the probability matrix to extract the characteristics of texture statistics. In this paper, first order statistical features like mean, variance, skewness, kurtosis, entropy and second order features like Angular Second Moment(energy), correlation, contrast and homogeneity [19] are selected.

D. Classification

Classification is a data mining (machine learning) technique used to predict group membership for data instances [14]. It is the problem of finding the model for class attribute as a function of the values of other attributes and predicting accurate class assignment for test data. It can be divided in two types: supervised and unsupervised. For the Classification in WEKA, we have supervised and unsupervised categories of classifiers. All the classifiers like lazy, tree, rules and nave comes under these categories only.The following classifiers are used here

The basic concept in SVM is the hyper plane classifier or linear separability. Two basic ideas are applied to achieve linear separability, SVM: margin maximization and kernels that is, mapping input space to a higher-dimension space (or feature space). SVM projects the input data into a kernel space. Then it builds a linear model in this kernel space. A classification SVM model attempts to separate the target classes with the widest possible margin [18]. A regression SVM model tries to find a continuous function such that maximum number of data points lie within an epsilon-wide tube around it. Different types of kernels and different kernel parameter choices can produce a variety of decision boundaries (classification) or function approximators (regression). In WEKA this classifier is called SMO.

The Multilayer Perceptron (MLP), a feed-forward back-propagation network, is the most frequently used neural network technique in pattern recognition [15]-[16]. Briefly, MLPs are supervised learning classifiers that consist of an input layer, an output layer, and one or more hidden layers that extract useful information during learning and assign modifiable weighting coefficients to components of the input layers. In the first (forward) pass, weights assigned to the input units and the nodes in the hidden layers and between the nodes in the hidden layer and the output, determine the output. The output is compared with the target output. An error signal is then back propagated and the connection weights are adjusted correspondingly. During training, MLPs construct a multidimensional space, defined by the activation of the hidden nodes, so that the two classes (malignant, and normal tissue) are as separable as possible.

The K algorithm can be defined as a method of cluster analysis which mainly aims at the partition of 'n'observation into 'k' clusters in which each observation belongs to the cluster with the nearest mean. We can describe K algorithm as an instance based learner which uses entropy as a distance measure. The benefits are that it provides a consistent approach to handling of real valued attributes, symbolic attributes and missing values. K is a simple, instance based classifier, similar to K Nearest Neighbour (K-NN). New data instances, x, are assigned to the class that occurs most frequently amongst the k-nearest data points, y_j, where j = 1, 2k. Entropic distance is then used to retrieve the most similar instances from the data set. By means of entropic distance as a metric has a number of benefits including handling of real valued attributes and missing values.

Statistical inferences are usually based on maximum likelihood estimation (MLE). MLE chooses the parameters that maximize the likelihood of the data, and is intuitively appealing. In MLE, parameters are assumed to be unknown but fixed, and are estimated with some confidence. In Bayesian statistics, the uncertainty about the unknown parameters is quantified using probability so that the unknown parameters are regarded as random variables. Bayesian inference is the process of analyzing statistical models with the incorporation of prior knowledge about the model or model parameters. Logistic regression, a special case of a generalized linear model, is appropriate for these data since the response variable is binomial. If you have some prior knowledge or some non-informative priors are available, you could specify the prior probability distributions for the model parameters. By Bayesťheorem, the joint posterior distribution of the model parameters is proportional to the product of the likelihood and priors.

IV. Experimental Results

The performance of the proposed method is evaluated using the dataset of 113 colon tissue images comprising 64 cancerous images and 49 normal images. To measure and investigate the performance on the selected classification methods namely MLP, SMO, K-star and Bayesian Logistic Regression we use the experiment procedures by WEKA. In Percentage split, 79-82% data is used for training and the remaining is for testing purposes.

In this study, all data is considered as instances and features in the data are known as attributes. Different performance matrix like TP rate,Precision, Recall and F-measure are presented in numeric value during training and testing phase. The summary of those results by running the techniques in WEKA is reported in Table I and III. Table II shows different types of error measurement like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE) and Root Relative Squared Error (RRSE).

TABLE I: Different Perfomance classifier in WEKA

Classifier	TPrate	Precision	Recall	F measure
BLR	0.81	0.838	0.81	0.816
K star	0.8	0.835	0.8	0.81
SMO	0.8	0.8	0.8	0.8
MLP	0.833	0.833	0.833	0.833

V. Result Analysis and Discussion

In this work, we examined the performance of different classification methods that could generate accuracy and some error to diagnosis the data set. According to Table I & III,

TABLE II: Error Measurement for different classifier in WEKA

Classifier	MAE	RMSE	RAE (%)	RSRAE (%)
BLR	0.2644	0.4364	54.263	89.5848
K star	0.2229	0.4453	45.7768	91.3613
SMO	0.2	0.4472	41.0811	91.7596
MLP	0.1905	0.4016	39.1667	82.2079

TABLE III: Perfomance Metrices in WEKA

Classifier	Accuracy (%)	Kappa statistics
BLR	80.9524	0.5758
K star	80	0.5294
SMO	80	0.4667
MLP	83.333	0.625

we can clearly see the highest accuracy and kappa statistics belongs to MLP compared with other classifiers. Based on the Kappa Statistic criteria, the accuracy of this classification purposes is substantial [17]. According to best average kappa statistic, the MLP classifier is best among others. Based on Table II, we can compare errors among different classifiers in WEKA. We clearly find out that MLP is the best, second best is the Bayesian Logistic Regression and K-star & SMO is moderate. The classifier which has a lower error rate will be preferred as it has, the more powerful classification capability and ability in terms of medical and Bioinformatics fields.

VI. Conclusion

Cancer is one of the most important health problems that threat the human life. Colon is one major constituent of the large intestine, and its cancer is a major reason of deaths in the western and industrialized world. A novel hybrid model is proposed, that combines both structural and statistical pattern recognition used for the detection of colon cancer and compare the classifiers using WEKA. The results, evaluates the four selected classification algorithms based on WEKA. The best algorithm in WEKA is MLP classifier with an accuracy of 83.333% and kappa statistics is 0.625. The MLP classifier which has a lower error rate will be preferred as it has the more powerful classification capability.

References

[1] R. S. Houlston, ”Molecular Pathology of Colorectal Cancer”, *J. Clinical Pathology*, vol. 54, pp. 206-214, Feb 2001.

[2] D. Altunbay et al., ”Color Graphs for Automated Cancer Diagnosis and Grading”, *IEEE Trans. Biomedical Eng.*, vol. 57, no. 3, pp. 665-674, Mar 2010.

[3] C. G. Demir et al., ”Automatic Segmentation of Colon Glands Using Object-Graphs”, *Medical Image Analysis*, vol. 14, pp. 1-12 , 2010.

[4] A. N. Esgiar et al., ”Microscopic Image Analysis for Quantitative Measurement and Feature Identification of Normal and Cancerous Colonic Mucosa”, *IEEE Trans. Information Technology in Biomedicine*, vol. 2, no. 3, pp.197-203, Sept. 1998.

[5] H. Kalkan, M.N.R. Duin and M. Loog, ”Automated Classification of Local Patches in Colon Histopathology”, *Proc. 21st Intl Conf. Pattern Recognition*, pp. 61-64, 2012.

[6] K. Masood et al.,”Co-Occurrence and Morphological Analysis for Colon Tissue Biopsy Classification”, *Proc. Fourth Int'l Workshop Frontiers of Information Technology*,2006.

[7] X. Li et al., ”Detection of Colon Cancer by Laser Induced Fluorescence and Raman Spectroscopy”, *Proc. IEEE Eng. In Medicine and Biology Ann.Conf.*, pp. 6961-6964, 2005.

[8] E. Ozdemir and C.G. Demir, ”A Hybrid Classification Model for Digital Pathology Using Structural and Statistical Pattern Recognition”, *IEEE Trans. Medical Imaging*, vol. 32, no. 2, pp. 474-483, Feb. 2013.

[9] Rathore et al. ”A Recent Survey on Colon Cancer Detection Techniques”, *IEEE/ACM Trans. Computational Biology and Bioinfomatics*, vol. 10, no. 3, May/June 2013.

[10] Cigdem Demir, S. Humayun gultekin, ”Learning the topological properties of brain tumors”, *IEEE/ACM Trans.on Computational biology and bioinformatics*, vol. 2, no. 3, July-Sep 2005.

[11] A. B Tosun and C. G demir, ”Unsupervised tissue image segmentation through object-oriented texture”, *Int Conf on Pattn Recog*, pp. 2517-2519, 2010.

[12] Hetal J. Vala and Astha Baxi, ”A Review on Otsu Image Segmentation Algorithm”, *Int J. of Advanced Research in Compt Science and Software Engg.*, vol. 2, no.3, pp. 2278-1323, Feb 2013.

[13] Navdeep Kaur and Usvir Kaur, ”Survey of pattern Recognition Methods”, *Int J. of Advanced Research in Compt Science and Software Engg.*, vol. 3, no. 2, Feb 2013.

[14] J. W. Han and M. Kanber, ”Data Mining Concept and Techniques”, *Morgan Kaufmann Publishers, Burlington*, Mar 2000.

[15] Duda, R. O. Hart and P. E., ”Pattern Classification and Scene Analysis”, *In: Wiley-Interscience Publication, New York*, 1973

[16] Bishop, C. M., ”Neural Networks for Pattern Recognition”, *Oxford University Press,New York*, 1999.

[17] I. H. Witten and E. Frank, ”Data Mining: Practical Machine Learning Tools and Techniques”, *2nd Edition, Morgan Kaufmann, San Francisco*, 2005.

[18] Yugal kumar and G. Sahoo, ”Study of Parametric Performance Evaluation of Machine Learning and Statistical Classifiers”, *Int.J. Information Tech and Compt Science*, no. 6, pp. 57-64, 2013.

[19] P. Mohanaiah, P. Sathyanarayana and L. GuruKumar, ”Image Texture Feature Extraction Using GLCM Approach”, *Int.J.of Scientific and Research Publications(IJSRP)*, vol.3, Issue 5, May 2013.