JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

AN EXPERIMENTAL STUDY ON THE MACHINE LEARNING TECHNIQUES FOR ORAL CANCER CLASSIFICATION

[1]Dr.K.Padmavathi, [2]Ms.C.Deepa
[1]Associate Professor, [2]Assistant Professor
[1]Department of Computer Science
[1]PSG College of Arts & Science, Coimbatore, India

Abstract : Oral Cancer is a considered as complex and wide spread cancer as it is rapidly evolving with a low median survival rate. Currently, duration and cost of the treatment process is long and very high due to its high recurrence and mortality rates. Accurate early diagnosis and prognosis prediction of cancer are become more essential to enhance the patient's survival rate using medical oncology. Using advanced technologies and machine learning techniques early detection of the deadly disease are made possible. Enabling automated detection and classification of the malignant lesions along computing the prognosis of the disease can be carried out using machine learning techniques with low cost and early diagnosis of the disease. In this paper, experimental study on machine learning technique for oral cancer classification has been carried on basis of defining the disease, diagnosis of the disease, classification of the disease in terms of stages of the lesion on basis of structure and finally prognosis and survival rate. Machine learning model is capable of learning the complex lesion patterns of the disease extracted from feature extraction and feature selection model. Classification of the patterns has been represented into stages. Classification results are highly discriminant with enhanced classification rate on the dynamic characteristics of the images. Evaluation of the technique is estimated using various datasets. The evaluation of the classification technique is based on the feature extraction and feature selection methods. Finally the performance analysis has done with respect to classification accuracy and execution time and attains the effective results on the cross fold validation of the dataset using confusion matrix on basis of precision, recall and f measure.

Index Terms **- Machine Learning, Oral Cancer, Classification, Staging, Feature Selection, Feature Extraction.**

I. INTRODUCTION

According to the American Cancer Society, Oral cancer is characterised by high mortality rate and morbidity around the world which occurs mostly in peoples over 40 years of age. Oral Cancer can be primarily found in the tissue of the mouth as dead cells. Oral Cancer can be diagnosed using several imaging methods like Computed Tomography or Magnetic Resonance Imaging. Acquired image using the imaging method has been employed to computer vision technique to determine the lesion as benign or malignant. Further processing is required to analyse the malignant lesion to classify it into several stages on basis of the structural characteristics and patterns of the lesion presence in the image [1].

Machine learning algorithm has been employed to classify the malignant lesion in addition to feature extraction and feature selection. In this paper, experimental study on machine learning technique for oral cancer classification has been carried using machine learning technique to classify the stage and predict the prognosis and survival rate of the patient with malignant lesion structures [3]. Analysis of the model is carried out on the highly discriminant patterns from the lesion extracted using feature extraction and feature selection model. In addition, capability of the learning model is computed using the complex lesion patterns of the disease. Further performance of the model is evaluated using various dataset [2].

Finally evaluation of the technique is estimated using various oral cancer datasets. The evaluation of the classification technique has been done in accordance to partitioning of the dataset into the training and testing data along the validation set. In addition, preprocessing model has been analysed to compute efficiency of the model on removing the noisy, containing some irrelevant or redundant information through preprocessing techniques [3]. The performance analysis has performed with respect to classification accuracy and execution time which is used to attain the effective results on the cross fold validation of the dataset using confusion matrix to compute the precision, recall and f measure to determine the accuracy and scalability of the models.

The rest of paper is structured as follows, section 2 describes the definition of the disease and staging of the disease on basis of clinical trials and pathology analysis, where section 3 describes the oral cancer dataset, while section 4 presents the types of the preprocessing and post processing techniques employed to the oral cancer dataset. In section 5, review of literature on machine learning techniques has been analysed with its advantages and limitation in depth and section 6 gives the brief overview of the classification techniques used in oral cancer. Finally Section 6 concludes the work.

II. Definition of the Important Terms

In this section, definition of the oral cancer and disease component has been provided with staging of the disease on basis of clinical trials and pathology analysis as preliminary step of the disease classification and prediction of prognosis on analysis of the dataset. They are:

2.1 Oral Cancer Definition

Cancer that forms in tissues of the oral cavity (the mouth) or the oropharynx (the part of the throat at the back of the mouth).It can be visualized as sore in mouth. Oral cancer includes cancers of lips, tongue, cheeks, hard and soft palates of the mouth.

2.2 Symptoms of the Oral Cancer

Symptoms of the Oral Cancer disease is swelling , thickening, lumps or bumps, rough spots, velvety white, red, speckled patches in mouth, unexplained bleeding and numbness in mouth, loss of sensation, pain in mouth , face and neck. Persistent sores in the face, neck and mouth which does not heal more than 2 weeks. In first stage, person has to difficulty in chewing and Swallowing. Finally Dramatic weight loss is occurred.

2.3 Screening Tools of the Oral Cancer

Screening of the cancer is carried out using Imaging tools like Histopathological Imaging and CT images.

2.4. Pathological Analysis of the Staging of the Disease

Staging of the disease is the basis of the lesion structure in the image. In this stage 0 is considered as abnormal, stage 1 is result of the tumor or lesion with size to be 2cm or less, stage2 will have size greater than 2 and smaller than 4cm, stage 3 is greater than 4 cm. Finally stage 4 is considers as tumor lesion spread to lymph node.

III. Dataset Description

The dataset used for oral cancer can be Histopathology or Computerized tomography or PET (positron emission tomography) is used to analysis.

3.1 Histopathology

Mendeley dataset contains images of oral lesions with histopathological results were collected from the archive of the department of tumour Pathology, which formed the initial source of dataset. The repository has 1224 images divided into two sets of images based on two different resolutions. Histopathology is used for diagnosing a disease based on the investigation of biological tissues, and to detect the presence of diseased cells in microscopic detail. It usually involves a biopsy.

3.2 Computerized Tomography

CT image is used to determine the stage 4 cancer to measure the cross sectional images to determine the structures like the hard palate or the upper or lower jawbone.

IV. Machine Learning Technique for Oral Cancer Classification

Oral Cancer Classification is carried out using the machine learning technique. In addition to processing of the image, preprocessing technique has to be employed to enhance the image quality and accuracy.

4.1. Analysis of pre-processing of oral cancer

Preprocessing of the image is carried out to remove the noise, contrast enhancement and normalization of the image to be classified. The Noise removal is carried out using selective median filter, contrast limited adaptive histogram equalization is employed to enhance the contrast of the image and normalization of the image. Feature extraction is employed to extract the feature of the image for the classification. Linear discriminant analysis and principle component analysis is employed to the extract the feature of the images.

4.2. Analysis of processing techniques of oral cancer

Oral cancer Classification is carried out on the extracted feature vectors containing the image patterns using following machine learning technique which is described as follows:

4.2.1. Decision Tree

A decision tree is a composed of classification and regression process as CART (Classification and Regression tree). A decision tree is a structured decision analysis. It is easy to learn and interpret. Decision tree indirectly performs selection of features. It works on both numerical and categorical data towards classification and prediction. Decision tree algorithms uses, Giri index, chi-square, information gain and reduction in variance of the features [4].

4.2.2. Random Forest

Random forest is the machine learning technique which can be used for classification and regression tasks of medical data. It works almost similarly as decision tree, it uses bagging method .Bagging is the combination of creating models and improve the output results[5]. Random forest combines two or more decision trees to predict the stages results on oral cancer disease. Therefore Random forest works by splitting a node as random subsets of the features.

4.2.3. Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the simple machines learning algorithm which produces accuracy with less computational power for cancer stage classification .SVM can be used for both classification and regression process on its main objective is to create classification models. It can be carried by identifying hyperplane in n number of features which classifies the data points. There can be many hyper-planes to differentiate data points. Hyper-planes are also called as decision boundaries. The

objects margins can be maximized using support vectors, by eliminating the support vectors the position and distances changes from the boundaries [6].

4.2.4. K-Nearest Neighbor (KNN)

KNN is simplest classification algorithm used in machine learning, it is suitable for both large and small datasets. It produces accurate results for more complex problems. KNN is used for classification and regression predictive models, which is mostly used for medical data classification. KNN is used Euclidean distance method to measure distance.. The distance measures are arranged in order to get the top most k-value and frequent class and then results in prediction output. KNN algorithm is also used for regression tasks by calculating averages of nearest objects in a class rather than calculating the mean object in a class [7].

4.2.4. Logistic Regression

Logistic regression is the machine learning algorithm for classification and prediction purpose. Logistic regression is classified into three types namely, Binary logistic regression, Multinomial logistic regression and Ordinal logistic regression. To predict the disease class on stages is set based on threshold value by estimated probability. Logistic regression is a straight technique; for binary/multivariate classification tasks [8].

V. Review of Literature for Oral Cancer Classification

In this section, various literatures employed for Classification of the oral cancer with stages has been examined in detail.

- B. Thomas et.al proposed ‘‘Texture analysis based segmentation and classification of oral cancer lesions in color images using ANN, “which employs Grey Level Co-occurrence Matrix (GLCM) and Grey Level Run-Length (GLRL) matrix are widely used for image characterization based on texture analysis.
  Selected texture discriminating features is used for classification of oral cancer lesions. Back propagation based Artificial Neural Network (BPANN) is used to compare and validate the performance of different feature sets.
  The classification accuracy is observed to improve with combination of GLCM, GLRL and intensity based first order features. Further improvement in accuracy is obtained by application of feature selection using boxplot analysis[9].

- A. Rana Et.al Proposed ‘‘Automated segmentation of gingival diseases from oral images” which identifies cause of tooth loss among people of all ages and are also correlated with systemic diseases.
  Machine learning classifier such KNN, trained with annotations from dental professionals that successfully provides pixel-wise inflammation segmentations of color-augmented intraoral images.
  The classifier successfully distinguishes between inflamed and healthy gingiva and its area under the receiver operating characteristic curve is 0.746, with precision and recall of 0.347 and 0.621 respectively [10].

- R. Girshick et.al propsoed ‘‘Rich feature hierarchies for accurate object detection and semantic segmentation”which typically combine multiple low-level image features with high-level context.
  It is a simple and scalable detection algorithm that improves mean average precision (mAP) by more than 30%.
  ANN is applied to localize and segment objects on training data is scarce and domain-specific fine-tuning yields a significant performance boost [11].

- X. Li, B. Aldridge et.al ‘‘Estimating the ground truth from multiple individual segmentations with application to skin lesion segmentation” employs a level-set based approach that solves the ground truth estimation in a probabilistic formulation.
  The prior pattern information is incorporated into the estimation model by adding a specially designed term in the energy function [12].

- W. Liuet.al proposed, ‘‘SSD: Single shot multibox detector” employs as method to detect the objects. SSD will be discretized the output space of bounding boxes into a set of default boxes over different aspect ratios and scales per feature map location.
  The network generates scores for the presence of each object category in each default box and produces adjustments to the box to better match the object shape.
  Additionally, the network combines predictions from multiple feature maps with different resolutions to naturally handle objects of various sizes [13].

VI. Tabular View of the Classification Techniques

The various classification techniques are used for classification. The following table shows the classification techniques used in oral cancer.

Table 1. Classification Techniques

S.No	Problem	Technique	Objective	Advantage	Disadvantage
1	Complex Structure of the Tumor	Back propagation based Artificial Neural Network (BPANN)	classification of oral cancer lesions	It identifies intensity based first order features	Computation time is high
2	Degradation of surrounding tooth structures, severe	Automated segmentation of gingival diseases	KNN with Annotation for Pixel Wise	The classifier successfully distinguishes between	Model is Complex to interpret as it affected by

	inflammation and gingival bleeding	from oral images	Segmentation	inflamed and healthy gingiva	irrelevant features.
3	low-level image features Processing	Artificial Neural Network	sliding-window detector	Domain-specific fine-tuning, yields a significant performance boost.	training data is scarce
4	Complex in ground truth estimation	level-set based approach	lesion segmentation	It capture pattern variation effectively	Localization of the segmentation leads to wron disease labels
5	Complex pattern extraction on various aspect ratio	SSD: Single shot multibox detector	Classification of resampling features	SSD model is simple and faster	Fails in feature resample with encapsulation

VII. Conclusion

An experimental study on machine learning algorithm for oral cancer classification has been carried out on image based dataset. Especially classification approaches analysed in this work is capable of the classifying the stages of the disease and predicting the survival rate of the patient. On analysis, it came to conclusion that stage of the disease and prognosis prediction has been carried out with high accuracy scalability. Finally experimental analysis of analysed technique to classify the disease into stages using machine learning model has proved on cross fold validation to demonstrate the effectiveness and robustness of the approaches.

VIII. Acknowledgment

The research is funded by Institutional Research Seed Grant, PSG College of Arts & Science, Coimbatore.

References

[1] Ahmad LG*, Eshlaghy AT, Poorebrahimi A, Ebrahimi M and Razavi AR, Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence, Health & Medical Informatics (2013), 2157-7420.

[2] Amy F. Ziober,Kirtesh R. Patel,Faizan Alawi, Phyllis Gimotty,4 Randall S. Weber, MichaelM. Feldman,Ara A. Chalian,Gregory S. Weinstein,Jennifer Hunt, and Barry L. Ziober, Identification of a Gene Signature for Rapid Screening of Oral Squamous Cell Carcinoma, American Association for Cancer (2018).

[3] Fatihah Mohd, Noor Maizura Mohamad Noor, Zainab Abu Bakar, Zainul Ahmad Rajion, Analysis of Oral Cancer Prediction using Features Selection with Machine Learning, ICIT 2015 The 7th International Conference on Information Technology.

[4] Girshick.R, J. Donahue, T. Darrell, and J. Malik, ‘‘Rich feature hierarchies for accurate object detection and semantic segmentation,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 580–587.

[5] Harikumar Rajaguru and Sunil Kumar Prabhakar, Performance Comparison of Oral Cancer Classification with Gaussian MixtureMeasures and Multi Layer Perceptron, The 16th International Conference on Biomedical Engineering p(2017) 123-129.

[6] Konstantina Kourou , Themis P. Exarchos , Konstantinos P. Exarchos ,Michalis V. Karamouzis , Dimitrios I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Computational and structural biotechnology journal, (2015),18-17.

[7] Marc Aubreville, Christian Knipfer, Nicolai Oetter, Christian Jaremenko, Erik Rodner, Joachim Denzler, Christopher Bohr, Helmut Neumann, Florian Stelzle, &Andreas Maier, Automatic Classification of Cancerous Tissue in Laserendomicroscopy Images of the Oral Cavity using Deep Learning, SCIENTIFIC Reports,(2017), 7: 11979.

[8] Martin Halicek, Guolan Lu, James V. Little, Xu Wang, Mihir Patel, Christopher C. Griffith, Mark W. El-Deiry, Amy Y. Chen, Baowei Fei, Deep convolutional neural networks for classifying head and neck cancer using hyperspectral imaging, J. Biomed. Opt. 22(6), 060503 (2017), doi: 10.1117/1.JBO.22.6.060503.

[9] Rana.A, G. Yauney, L. C. Wong, O. Gupta, A. Muftu, and P. Shah, ‘‘Automated segmentation of gingival diseases from oral images,’’ in Proc. IEEE Healthcare Innov. Point Care Technol. (HI-POCT), Nov. 2017, pp. 144–147.

[10] Thomas.B, V. Kumar, and S. Saini, ‘‘Texture analysis based segmentation and classification of oral cancer lesions in color images using ANN,’’ in Proc. IEEE Int. Conf. Signal Process., Comput. Control (ISPCC), Sep. 2013, pp. 1–5.

[11] X. Li, B. Aldridge, J. Rees, and R. Fisher, ‘‘Estimating the ground truth from multiple individual segmentations with application to skin lesion segmentation,’’ in Proc. Med. Image Understand. Anal. Conf., vol. 1, London, U.K., 2010, pp. 101–106.

[12] W. Liu, D. Anguelov, D. Erhan, and C. Szegedy, ‘‘SSD: Single shot multibox detector,’’ in Proc. Eur. Conf. Comput. Vis., 2016, pp. 21–37.