International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 12, Number 5 (2019), pp. 596-601

Oral Cancer Analysis Using Machine Learning Techniques

Lavanya L[1], Dr. Chandra J.[2]

[1]*PG Scholar, Department of Computer Science, CHRIST (Deemed to be University), India.*

[2]*Associate Professor, Department of Computer Science, CHRIST (Deemed to be University), India.*

ORCIDs: 0000-0002-9259-9377(Lavanya), 0000-0001-8324-8746 (Dr. Chandra)

Abstract:

Oral cancer staging is most required task of examining treatment and required medication for the patients. In oral cancer staging is of two types, namely, clinical and pathological. TNM (Tumor, Node, and Metastasis) staging is clinical system of predicting oral cancer stages, on the other hand Histology, p63 and podoplanin expressions are pathological staging system which is obtained after biopsy test. These staging systems are used in machine leaning techniques to analyze the different stages of oral cancer. The main aim of the paper is to classify different stages of oral cancer using machine learning techniques. The experimental work is based on clinical and pathological staging system. The data set used for this research work is based on Oral Leukoplakia. The data transformation is applied to standardize the data and the features were extracted using correlation coefficient. The extracted features were classified using Decision tree and random forest which are compared against other popular classification methods like SVM, KNN, MLP and Logistic Regression. From the experimental work, it is found that the various stage classification of oral cancer can be classified efficiently with help of Random Forest and Decision Tree. So the classification of various oral cancers can be performed with help of random forest and Decision Tree. .

Keywords: Machine Learning, Decision tree, SVM, KNN, Random Forest, oral cancer, TNM, OSCC, Prediction.

I. INTRODUCTION

Oral cancer is stated as the out of control in growth of cells that secure and cause damage to surrounding tissue [14]. Oral cancer finds fewer dead cells in the tissues of mouth at the beginning of oral cancer development which is known as a lesion. Where metastasis means, dead cells present in the faraway site of the affected area or internal parts of the body. There are various types of cancer out f those squamous cell carcinomas occurrence is 90% in the medical field which is also called as OSCC (Oral Squamous Cell Carcinomas) [15]. The Bio-specimens and clinical profiling of a similar tumor and lesion-free specimens can be detected in various parts of the body by expression patterns and classifying cancer-free from OSCC tissues. Machine Learning algorithms are utilized to predict the different bio-specimens for OSCC which would classify cancer-free specimens and lesion specimens, Later which are analyzed for the staging of oral cancer. Predictors will find the accuracy on cross-validation, by utilizing three validation test sets and various levels in cancer. Identification and validation of specimens would predict various tumor sizes and occurrence of the lesion in tissues, which helps in predicting different stages of oral cancer, [6].The main aim of the current process is towards developing a new Predictor tool to predict the stages of tumor growth in oral cancer. Oral cancer arises in areas like the front side of the tongue, top and bottom of the mouth (below the tongue), insides of the cheeks and lips, gum and area behind the wisdom teeth. Symptoms of oral cancer are, The most common sign of cancer is a sore or ulcer that does not heal, and may cause pain or bleeding, White or red sores in the mouth, Lips, gums, or tongue that do not heal, A lump or mass in the mouth, Loose teeth, Trouble chewing or swallowing, Jaw inflammation, Difficulty speaking and Chronic sore throat [20].

The risk factors of oral cancer consist of certain actions such as, tobacco chewing and alcohol drinking which are considered as the major risk of oral cancer. In India, consumption of beetle nut is common which also affects the inner cheek area of the tooth [21]. Other risk factors such as Human Papillomavirus (HPV), Age, Gender (Lip and oral cavity cancer are more common in men than in women).

Now coming to oral cancer staging which depends on the affected area in the mouth, size of a tumor and the presence of dead cells in lymph nodes or inner parts of the mouth. The different stagings in oral cancer are as follows [22]

- Stage 0- Abnormal (no damage cells present in the outer layer of the tissue).
- Stage 1- Tumor size is said to be 2cm or less.
- Stage 2- Tumor size is larger than 2cm and smaller than 4cm.
- Stage 3- Tumor size is larger than 4cm or not larger than 3cm. (dead cells increase in the lymph node on the same side of the neck).
- Stage 4a- Tumor size larger than 3 cm (growth in the lymph node on the same side of the neck), but no larger than 6 cm (multiple lymph nodes on the same side of the neck), larger than 6 cm (lymph nodes on opposite side or both sides of the neck).
- Stages 4b- Tumor cells have increased to lymph node that is larger than 6 cm.
- Stages 4c- Tumor cells have increased to different organs of the body such as lungs, bones, liver.

Machine learning is an area of artificial intelligence that uses statistical methods to provide computer systems with the ability to "learn" from expertise. Machine learning algorithms use process techniques to be told information's directly from knowledge while not counting on a planned equation as a model. Machine learning is of two types such as, supervised learning and unsupervised learning, where supervised learning trains a model on known input and output data which can predict future outcomes. Whereas unsupervised learning identifies hidden patterns or structures in input data.

Supervised machine learning builds a model that produces predictions based on substantiation in the presence of apprehension. A supervised learning algorithm takes a collection of data and actions to produce outcome data and prepares a model to generate predictions for the response to new data. The Supervised method uses classification and regression techniques to develop a predictive model. Regression method predicts continuous actions whereas Classification techniques predict individual action. The common classification algorithms are support vector machine (SVM), decision trees, K-nearest neighbour, Naïve Bayes, logistic regression, and neural networks [13].

II. RELATED WORK

Fatihah Mohd et al. researcher objective was to predict the primary stage of oral cancer with accurate results using less attributes by using Naïve Bayes, Multilayer Perceptron, K-Nearest Neighbors and Support Vector Machine methods they resulted in oral cancer stage and analyzed an increase in classification accuracy. [4].

Ahmad LG et al. researcher objective was to develop Models for medical practitioners. By using Decision Tree, Support Vector Machine, Artificial Neural Network methods and analyzed the accuracy of DT, ANN and SVM which are high. With highest accuracy and least error rate SVM classification model is best for predicting breast cancer recurrence. When compared to ANN and DT, The results prove that SVM are the better classifier for prediction [1].

Harikumar Rajaguru and Sunil Kumar Prabhakar researcher objective was to compare the classification accuracy of the TNM staging system using Multi Layer Perceptron (MLP) and Gaussian Mixture Model classifiers. Comparison of both classifiers here provided a better result as average accuracy for the stages. Extreme Learning Machines (ELM) was used as a post classifier later for the oral cancer analysis and the performance of ELM classifier was compared with performance of both GMM and MLP [5].

Amy F. Ziober et al. used SVM classifier method to detect OSCC tumors by examining expression profiling on patient and extracting RNA plus microarray analysis a gene expression signature predicts OSCC tissue from normal[2].

Marc Aubreville et al. objective was to evaluate a novel automatic approach for OSCC diagnosis by using deep learning and CNN methods on CLE images. Here CNN method was to find patch-extraction of images, training the data and classifying [10].

Shreyansh A et al. used dataset consisting of 251 RVG X-rays images which was later divided into test and train sub-datasets for experimenting different models such as deep learning, ANN, transfer learning and CNN. Hence they achieved accuracy of 88.46% overall [19].

Martin Halicek et al. the researcher experimented on OSCC, Thyroid cancer and head and neck sample tissues using CNN classifier to identify cancer. The result was CNN produced 80% accuracy in detecting cancer [11].

Ramzi Ben Ali et al. the aim of researcher was to classify dental X-rays images into decayed or normal tooth images and develop a new model to find dental issues in X-ray images using deep neural network technology [17].

Konstantina Kourou et al. the researchers produced a comparative study of various machine learning applications in different types of cancer prognosis and prediction. The study used heterogeneous data for developing models using machine learning techniques such as ANN,BN,SVM and decision tree by feature selecting and classification methods[8].

Wafaa K. Shams and Zaw Z. Htike researchers' objective was to predict oral cancer development in OPL patients, where machine learning techniques were used with the help of gene expression profiling. The researcher used SVM, MLP, Regularized least squares (RLS) and deep neural networks for investigating the oral cancer development in OPL patient's records [25].

From the related work, it is observed that methods and materials used previously are included mainly on detecting cancer presence, classification of cancer types and the comparison of various machine learning algorithms. Hence, staging of oral cancer is an importance task in the oral cancer diagnosis. This work is not done by any researcher, which is a most required task in examining prognosis as well as treatment of cancer patient for medical practitioner. Thus current study invokes on applying various supervised machine learning methods which are focused on analyzing efficient staging in oral cancer development.

III. METHODOLOGY

In this section a detailed description about dataset and machine learning algorithms used in this study. In order to start the oral cancer stage prediction process, it is required to know more about medical terms and procedures from dental doctors, therefore a discussion was done with few dentists for the clarity of oral cancer concepts.

A. Decision Tree

A decision tree is a structural diagram used to produce solutions to a problem based on certain conditions. Decision Tree is mostly used in classification problems which are the part of supervised machine learning algorithm. A tree has many scenarios in life and which is also included in machine

learning in deep by covering both classification and regression trees also known as CART(Classification and Regression tree). A decision tree is a structured flowchart. Where each internal node indicates a test on attribute each branch represent an output of test and each leaf of end node consists a class label. The root node is the top most nodes in a tree.

To represent decisions and decision making, decision tree can be used in decision analysis. As the name indicates it uses a tree like model of decision.

Pros of Decision tree, It is easy to learn ad create, interpret and view. Decision tree indirectly performs selection of features. It works on both numerical and categorical data which can also handle harder problems.

Cons of Decision tree, creating over complex trees do not generalized well in decision tree. It is also known as over fitting. Decision tree can be unbalanced because small variants in information may result in a complete tree generating. It consumes more memory space. Decision tree is little difficult for preparing data.

The common Decision tree algorithms used are, Giri index, chi-square, information gain and reduction in variance [3].

B. Random Forest

Random forest is simple and stable algorithm in machine learning. It is also most commonly used algorithm which produces accurate results. Random forest can be used for classification and regression tasks. It is a supervised machine learning algorithm. One major preferred standpoint of random forest is, that it tends to be utilized for both classification and regression problems, which frame the greater part of current machine learning systems. It works almost similarly as decision tree, it uses bagging method .Bagging is the combination of creating models and improve the output results. Random forest combines two or more decision trees to predict the outcome results. It adds randomness for the trees instead of seeking for features while sub-dividing the nodes. It finds the best attribute among the features of random subset. It could result in better model at the end. Therefore Random forest works by splitting a node as random subsets of the features. Another advantage of random forest is that it is easy to calculate the importance of close features on prediction process [18].

C. Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the simple machine learning algorithm which produces accuracy with less computational power.SVM can be used for both classification and regression process. But its main objective is to create classification models. It can be done by identifying hyper-plane in n number of features which classifies the data points. There can be many hyper-planes to differentiate data points. The attributes which is found on either side of the hyper-planes by data points are of different classes. Hyper-planes are also called as decision boundaries. Data points are nothing but a support margin which are closer to the boundaries and includes the position and distance. For faraway objects margins can be maximized using support vectors, by eliminating the support vectors the position and distances changes from the boundaries. These above studies help to build SVM model [23].

D. K-Nearest Neighbor (KNN)

KNN is simplest classification algorithm used in machine learning, it is suitable for both large and small datasets, though it is simplest algorithm It produces accurate results for more complex problems. KNN is used for classification and regression predictive models, which is mostly use in industry for classification issues. KNN considers three aspects such as, ease to interpret; calculation time and prediction power.KNN algorithm is commonly used for interpreting and low calculation time. In KNN firstly the class is divided by boundary then identifies the mean of each class to calculate the distance between mean object and other objects present in the class. Suppose the distance is longer and near to mean class of neighbor then the object could belong to neighbor class. To calculate the distance between objects KNN shall use various distance measures. The commonly used distance measure in KNN is Euclidean distance method. 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].

E. Logistic Regression

Logistic regression is the most celebrated machine learning calculation after linear regression. From multiple points of view, linear regression and logistic regression are comparative. Be that as it may, the greatest contrast lies in what they are utilized for. Linear regression algorithms are utilized to predict/forecast values but logistic regression is used for classification tasks.

Logistic regression is of three types namely, Binary logistic regression, Multinomial logistic regression and Ordinal logistic regression. To predict the data class a threshold is set based on threshold value the classes are classified by estimated probability. Decision boundary can either be linear or non-linear to make decision boundary more complex, polynomial order is increased.

Logistic regression is a straight technique; however the expectations are changed utilizing the calculated capacity. Logistic regression is a straightforward calculation that can be utilized for binary/multivariate classification tasks [9].

F. Multi Layer Perceptron (MLP)

A multilayer perceptron (MLP) is one of the methods in artificial neural network that produces a set of outputs from a set of inputs. An MLP is multiple layers of input nodes joined as a directed graph between the input and output layers, where the nodes travel on single path. Using back propagation method predictive model are developed by training the dataset. MLP is a supervised learning of deep learning method. Thus there are multiple neurons layers, it is deep learning technique. MLP is mostly used for supervised models and neuroscience and as well for parallel distribution processing MLP is commonly used in speech recognition, image recognition and translation of machine languages [12].

IV. RESULT AND DISCUSSION

A. Dataset description

Prediction model of oral cancer were developed using Oral Leukoplakia dataset form National Library of Medical (Nation center for biotechnology information) of U.S for the year 2011[15].Since Indian dataset was not provided because of certain rules and regulations by Indian medical government. Oral Leukoplakia dataset is a pathological data which is produced after biopsy tests. This dataset was very easy in processing since it was a preprocessed dataset for the repository directly. Though it was preprocessed dataset it was clear to select the features for prediction process. The dataset contained 12 attributes namely, age, gender, race, Alcohol habits, Smoking habits, histology at baseline, histology(breakdown), p63 expression, podoplanin expression, treatment, oral cancer free survival rate, cancer outcome and time of biopsy.

Histology is a diagnosis test results namely, hyperplasia and dysplasia. A result of reaction to a bacterial throat infection can be viewed when there is improvement of lymph nodes in the neck due to hyperplasia. While hyperplasia would not find cancer growth the other parts of the body. Dysplasia is the next level of Hyperplasia. Dysplasia contains Mild, Moderate and Sever levels which symbolizes the staging of oral cancer [24].P63 and Podoplanin expression are the methods taken up to test the presence of dead cells in tissues.

The Oral Leukoplakia dataset contained 12 attributes namely, age, gender, race, Alcohol habits, Smoking habits, histology at baseline, histology(breakdown), p63 expression, podoplanin expression, treatment, oral cancer free survival rate, cancer outcome and time of biopsy. Out of those 12 attributes six attributes namely Alcohol habits, Smoking habits, histology (breakdown), p63 expression, podoplanin expression, cancer outcome were used for prediction process. These attributes were selected using correlation of feature selection method. The featured attributes where used to train a model by supervised algorithms to predict the result.

The result was achieved by developing a predictor tool which was web-based tool to predict TNM and Pathologic staging. The predictor tool was developed using Machine Learning algorithms such as Decision tree, SVM,KNN, Naive Bayes, Random forest, Logistic regression, Multi layer perceptron (MLP) by coding them in python language.

Hence, the prediction result of Oral Leukoplakia dataset is shown below Figure 2.

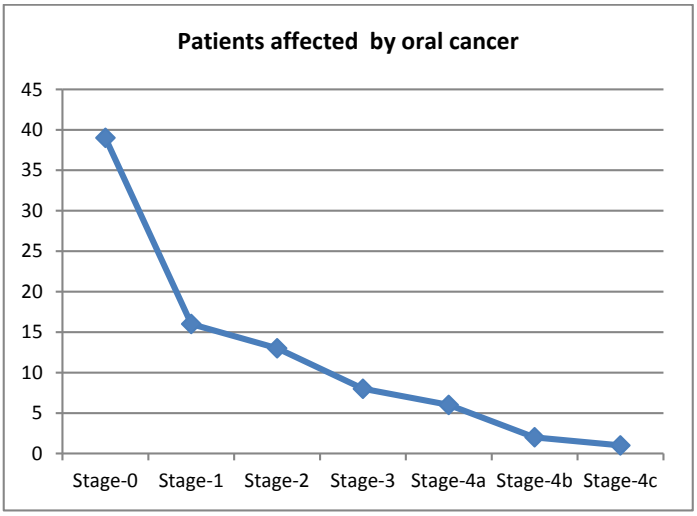

Figure 1 Prediction result of dataset.

In Fig 2.the graphical representation shows the prediction results of oral cancer patients affected in different stages are produced. This result is based on Oral Leukoplakia dataset. Hence 39 patients fall under 'stage-0' who all can be healed or there is no oral cancer development. Similarly 16 records in stage-1, 13 records in stage-2, 8 records in stage-3, 6 records in stage-4a, 2 records in stage-4b and 1 record in stage-4c.

The prediction result was analyzed for accuracy results of various machine learning algorithms, by applying three layer of cross-validation on accuracy was calculated. Table 1 justifies the three folds of cross-validation results in percentage. The below fig 2 represents the accuracy result of various machine learning algorithms.

Table I: 3-Folds of Cross-Validation Of Different Machine Learning Algorithms

Classifiers	Fold-1(%)	Fold-2(%)	Fold-3(%)
Decision Tree	82.759	82.759	83.703
SVM	82.759	82.759	82.553
KNN	82.759	82.759	80.255
Random Forest	79.310	79.310	81.404
Logistic reg.	68.966	68.966	59.236
MLP	75.862	75.862	69.704

The table 1 represents the percentage of 3 folds of cross-validations. This helps to calculate the accuracy rate of various algorithms.

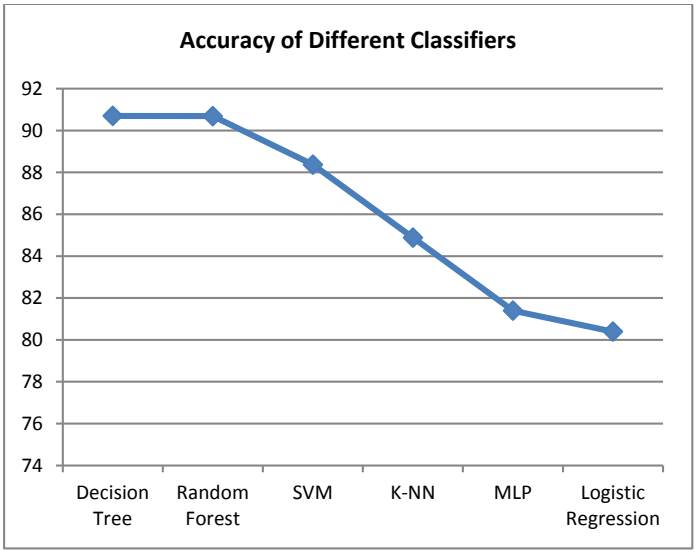

Figure 2. Accuracy results of different classifiers

The Fig 2 represents the graphical representation of various accuracy rates of different machine learning algorithms, in which the Decision tree is considered as 90.688%, Random Forest is 91%, SVM is 88%, KNN is 85%, MLP is 81% and Logistic Regression gives 80% of accuracy. It is clear that Decision tree and Random Forest algorithms produce same accuracy rate.

V. CONCLUSION

In diagnosis of Oral Cancer, the staging is one of the important tasks to be performed by medical practitioners of the cancer field. Hence it is important to classify different stages in Oral Cancer to give effective treatment for the cancer patient. The data transformation is applied to standardize the data and the features were extracted using correlation coefficient. The extracted features were classified using Decision tree and random forest which are compared against other popular classification methods like SVM, KNN, MLP and Logistic Regression. The current work implies the prediction of different stages in oral cancer and comparing the accuracy of various machine learning techniques by cross-validation. The result of two algorithms Decision tree and random forest gives the better accuracy results.

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] Decision tree, Swapnil Yeolekar, https://www.quora.com/Can-you-explain-a-decision-tree-in-simple-terms(2017).

[4] 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.

[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] Head-and-Neck-squamous-cell-carcinoma,https://en.wikipedia.org/wiki/Head_and_neck_squamous-cell_carcinom, Wikipedia(2018).

[7] K-Nearest Neighbors, Tavish Srivastava, https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

[8] 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.

[9] Logistic Regression, https://hackernoon.com/introduction-to-machine-learning-algorithms-logistic-regression-cbdd82d81a36

[10] 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.

[11] 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.

[12] Multi Layer Perception, https://www.techopedia.com/definition/20879/multilayer-perceptron-mlp Machine Learning, https://www.mathworks.com/discovery/machine-learning.html

[13] Oral cancer, https://www.webmd.com/oral-health/guide/oral-cancer#1

[14] Oral cancer, https://en.wikipedia.org/wiki/Oral_cancer

[15] Oral Leukplakia LM328"https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM652761

[16] Ramzi Ben Ali, Ridha Ejbali and Mourad Zaied, Detection and Classification of Dental Caries in X-ray Images Using Deep Neural Networks, The Eleventh International Conference on Software Engineering Advances,(2016), 978-1-61208-498-5.

[17] Random forest, https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[18] Shreyansh A. Prajapati, R. Nagaraj and Suman Mitra, Classification of Dental Diseases Using CNN and Transfer Learning, 5th International Symposium on Computational and Business Intelligence, (2017), 978-1-5386-1772-4/17.

[19] Symtoms of oral cancer, https://www.ahns.info /resources/education/patient_education/oralcavity/]

[20] Squamous-cell-carcinoma-cancer,https://en.wikipedia.org/wiki/Head_and_neck_squamous-cell_carcinoma

[21] Staging of Lip and Oral Cavity Cancer , https://www.ahns.info/patient-information/

[22] Support vector machine, Rohith Gandhi, https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (2018)

[23] The biopsy report: A Patient's Guide, https://oralcancerfoundation.org/discovery-diagnosis/biopsy-report-patients-guide/

[24] Wafaa K. Shams and Zaw Z. Htike, Oral Cancer Prediction Using Gene Expression Profiling and Machine Learning, International Journal of Applied Engineering Research, (2017) ,0973-4562.