A Review on Early Detection of Oral Cancer using ML Techniques

Sandhya N. Dhage
Assistant Professor, Computer Science & Engineering, G. H. Raisoni Academy of Engineering & Technology, Nagpur, India

Abstract - Discovering cancer in an early stage leads to an early treatment which lowers the risk of morbidity and mortality. The diagnosis of oral cancer remains a challenge to dental profession particularly in the detection, evaluation and management of early phase oral cancer. Due to lack of timely diagnosis using conventional methods, machine learning approach is used for detecting and classifying oral cancer at an earlier stage. Machine learning techniques are used to model the progression and treatment of cancerous detection and can predict future outcomes of cancer type effectively. Combination of efficient machine learning and feature selection algorithms gives better results in early diagnosis and prognosis of oral cancer. The main goal and contribution of this review paper is to summarize the use of machine learning techniques for accurate prediction of oral cancer at an earlier stage.

Keywords: Early prediction, Oral cancer, Machine learning.

I. INTRODUCTION

Oral cancer is the sixth common malignancy and becomes the major cause of cancer morbidity and mortality worldwide as more deaths occur every year from oral cancer. Maharashtra has the highest incidence of mouth cancer in the world. Clinical diagnosis for early detection of cancer leads to an early treatment, which lowers the risk of morbidity and mortality. Screening program approach are implemented by clinical doctor improves the chances of discovering cancer early increases so that people with cancer can get an early treatment. Clinical diagnosis collects data and features from patient's history that leads to problem in the diagnosis because most of the disease shares the same clinical features and scaling. Advanced clinical practices such as surgery, radiation and chemotherapy are used for treatment of oral cancer but the mortality rate associated with oral cancer has increased in the last 40 years. Tumors can be benign, premalignant or malignant. Malignant tumors are cancerous. Most of the people die as a result of the malignancy in oral cancer. Early evaluation of oral precancerous lesions is useful in reduction of oral cancer mortality rates. The diagnosis of Oral cancer remains a challenge to the dental profession, particularly in the detection, evaluation and management of early phase alterations or frank disease Prediction of Oral Leukoplakia (premalignant) and Oral Squamous Cell Carcinoma becomes a challenging task. The accurate prediction of a disease outcome is one of the most interesting and challenging tasks for physicians.

Because of difficulty to diagnose clinical diseases, many experts studied the solutions from both the medicine and computer science. Many scientists applied different methods, such as screening in early stage and developed new strategies for the early prediction of cancer treatment outcome. In the field of medicine, advanced technologies are adopted and large amounts of cancer data have been collected and are available to the medical research community. In the present era, machine learning methods have become a popular tool for medical researchers. Variety of machine learning methods such as feature selection and classifications are widely applied in the diagnosis of cancer. Machine learning techniques are used to discover and identify patterns and relationships from complex datasets. For efficient diagnosis of oral cancer, most relevant features can be selected using efficient feature selection method to achieve classification with higher accuracy based on efficient classification method. The paper is organized as follows: Section II describes the overview of oral cancer. Section III provides the review of studies that make use of machine learning methods regarding the detection of oral cancer. Section IV discusses the machine learning techniques. Section V concludes the manuscript.

II. ORAL CANCER

Oral cancer develops in the squamous cells found in mouth, tongue, and lips and belongs to a larger group of cancers called head and neck cancers. Mostly oral cancers are discovered after they have spread to the lymph nodes of the neck. Types of oral cancers are lips, tongue, inner lining of the cheek, gums, floor of the mouth, hard and soft palate. The symptoms for an oral cancer that raise the suspicion of cancer and needs proper treatment at an earlier stage are: 1) Patches inside the mouth or on lips that are white, red or mixture of white and red, 2) Bleeding in the mouth 3) Difficulty or pain when swallowing, 4) A lump in the neck. Treatments for Oral Cancer are not successful which include surgery, radiation therapy and chemotherapy because 70% of the cases relapses and the results in death. If the lesion is not diagnosed early, treatment becomes unsuccessful because many times, it is ignored and the patient reports late when the lesion is untreatable.[2]

Oral cavity cancer (OC) is the sixth to eight most common cancer around the world and is a major health concern

over the world called as a malignant neoplasm on the lip or in the mouth.[2] The pre malignant lesions occurred at the previous situation and clinical screening methods are used to note morphologically altered tissue in which cancer is more likely to occur than in normal tissue and such lesions may exhibit epithelial dysplasia (ED) on histopathologic examination. When lesions are easiest to remove and most likely to be cured, screening techniques are used to detect mouth cancer or precancerous lesions that may lead to mouth cancer at an early stage. Screening techniques include Vital Staining, Light-based detection systems, Histological Techniques, Imaging diagnostic techniques, Cytological Techniques, Molecular Analyses, Imaging diagnostic techniques, *Onco-chip*. But screening methods have not proved to be successful to save lives, so clinicians faces the challenges in the oral exam for oral cancer screening

Early detection and diagnosis of oral cancer can improve patient survival and reduce morbidity rate. Therefore new computer science methods are currently applied for accurate diagnosis.

III. LITERATURE SURVEY

In the paper [1], a technique is proposed to detect cancers present in mouth provided by an Orthopantomogram. A novel mathematical morphological watershed algorithm is proposed to preserve these edge details as well as prominent watershed on images leads to over segmentation even though it is pre-processed. To avoid over segmentation, marker controlled watershed segmentation is used to segment tumours.

In the paper[2], hybrid model is proposed which consist of two stages, where ReliefF-GA feature selection method to find an optimal feature of subset is used in first stage and ANFIS classification is used to classify survival of patient after certain years of diagnosis. The proposed prognostic model was experimented on two groups of oral cancer dataset which consist of clinicopathologic markers and genomic markers. It is experimented that that the proposed model is more accurate with the use of both types of dataset and the other methods of artificial neural network, support vector machine and logistic regression. This prognostic model can be used to help clinicians in the decision support stage and to identify the high risk markers to better predict the survival rate for each oral cancer patient.

Wafaa K. ShamsandZaw Z. Htike[3], predict the possibility of oral cancer development in OPL patients. They have used Fisher discriminate analysis to select relevant features from the gene expression array. Support vector machine (SVM), Regularized Least Squares (RLS), multi-layer perceptron (MLP) with back propagation and deep neural network (DNN) are used as classifier techniques.

Konstantina Kourou, Themis P. Exarchos[4], presented a review of recent ML approaches employed in the modelling of cancer progression. In the paper, different predictive models are discussed based on various supervised ML techniques as well as on different input features and data samples.

K. Anuradha1 and K.Sankaranarayanan[6] has done detailed survey on various methods analyzed by the researchers for the detection of oral cancers at an earlier stage. Various methods for the identification and classification of cancers are compared. Each step of the cancer detection algorithms is given.

Shikha Agrawal, Jitendra Agrawal [7] has given the survey of various neural network technologies for classification of cancer which is the burning research area in medical science.

Hakan Wieslander, Gustav Forsli, Ewert Bengtsson[8],have proposed that Convolutional Neural Networks have been proven to be accurate for image classification tasks. The performance was evaluated for two different network architectures, ResNetand VGG by using two datasets containing oral cells and cervical cells. The results shows that ResNet was prereferable network,with a higher accuracy and a smaller standard deviation.

Neha Sharma and Hari Om [9] proposed ED&P framework which is used to develop a data mining model for early detection and prevention of malignancy of oral cavity.

K. Anuradha and Dr. K. Sankaranarayanan[10] presented the work to detect oral cancers using image processing. Linear contrast stretching is used to remove noise from the dental X – Ray Image which is used as input. Marker controlled watershed segmentation is improved and used to segment tumors from the enhanced Image. The segmentation algorithms are compared for speed and accuracy and found that improved algorithm provides better segmentation.

Fatihah Mohd, Noor Maizura, Mohamad Noor[11] had worked on the integrated diagnostic model with hybrid features selection methods for diagnosis of oral cancer that determine the attributes reduces the number of features that are collected from a variety of patient records. Updatable Naïve Bayes, Multilayer Perceptron, K-Nearest Neighbors and Support Vector Machine classifiers are used to predict the diagnosis of patients with oral cancer. Further they have added that the support Vector Machine outperforms other machine learning algorithms after incorporating feature subset selection with SMOTE at pre-processing phases.

IV. MACHINE LEARNING

The real world problems are solved by machine learning which build a model that is good and useful approximation to the data. Machine learning is popular today because of growing volumes and varieties of available data, cheaper computational processing and more powerful, and affordable data storage. Machine learning produces models quickly and automatically that can analyze bigger, more complex data and deliver faster, more accurate results even on a very large scale. Machine learning model gives predictions with high accuracy used to take better decisions and smart actions in real time without human intervention. It is the need to develop newer algorithms to advance the science of machine learning and use its application in variety of research areas including health care. Current study has revealed that cancer can be cured with the use of machine learning. The use of machine learning and AI tools in basic and translational cancer research mark the beginning of a new era for personalized medicine, characterized by quick and advanced data analysis, which was previously unattainable. vast amounts of data along with machine learning algorithms helps to fight with cancer in many ways such as diagnosis, treatment, and prognosis of cancer. Machine learning helps to customize the therapy according to the patient, which is not possible otherwise. Given the enormous amounts of Electronic Medical Records (EMR) which is generated and recorded by various hospitals, it is possible to use 'labeled' data in diagnosing cancer. The different types of machine learning algorithms are applied on EMR databases and finds hidden patterns which helps in diagnosing cancers. Natural Language Programming (NLP) are used for analyzing doctor's prescriptions and deep learning neural networks are deployed to analyze CT and MRI scans. Diagnosis can be done with big data and machine Learning. The diagnosis is accurate when it depends on the amount and quality of the data available. If dataset is large, machine learning algorithm will be able to query the database which find similarities and produce accurate predicted models.

Big data and Machine learning are used widely for diagnosis and treatment of diseases. Big data and machine learning tools are used for customising treatments depending on the patient's history, their gene sequence, results of diagnostic tests, a mutation found in their genes or a combination of their genes and environment. Big data and Machine learning also help to revolutionise drug discovery. Despite of Big data and machine learning improves the process of diagnosis, treatment and drug discovery in treating cancer, scientists faces many challenges in the area. Since data is not digitized in all hospitals, it is captured and recorded using old methods and cannot be processed using advanced technologies.

Machine model is based on learning process which is divided into two training and testing process. In training process, learning algorithm is used in which features are learned from input samples in training data and build the learning model. In the testing process, production data is tested by learning model which uses the execution engine to make the prediction. The result of learning model is tagged data which gives the final prediction or classified data.

V. MACHINE LEARNING TECHNIQUES

Machine learning techniques are classified into three broad categories as follows:

A. Supervised learning

In supervised learning labeled examples are given as an input which are trained and the desired output is known. Training dataset consist of both features and labels. The task of supervised learning is to infer a function from labeled training data consisting of a set of training examples. It is used to construct an learning model which predicts the label of an object given the set of features. In supervised learning, the learning algorithm takes a set of features as inputs along with the corresponding correct outputs, and learning is performed by algorithm after comparing its actual output with correct outputs. If error occurs after comparison, it then modifies the model accordingly. It training data is missing; model is not capable to infer prediction correctly. Supervised learning is used in applications where classification is done on some data and to predict some data. For example, classifying whether a patient has disease or not.

Tasks of supervised learning are divided into two categories as classification and regression. In classification, the label is discrete, while in regression, the label is continuous. For example, astronomy problem is classification problem in which it detects the object and classifies it in distinct categories as a star, a galaxy. On the other hand, finding the age of an object based on observations is regression problem where the label (age) is a continuous quantity.

B. Unsupervised Learning

Unsupervised learning used unlabeled data as input and learning is done to explore the data and find similarities between the objects. It discovers the labels from the data itself. Unsupervised learning is used in applications where transactional data is given. For example, identify customers with similar attributes and grouped them so that they can be treated similarly in marketing campaigns. In supervised learning, task is performed to classify object in galaxy and star. On the other hand, unsupervised learning is applied when detailed observations of distant galaxies are given and it determines which features or

combinations of features are most important in distinguishing between galaxies.

Clustering is unsupervised task where a set of inputs is divided into groups whereas in classification, the groups are not known before. Self-organizing maps, nearest-neighbor mapping, k-means clustering and singular value decomposition are popular unsupervised techniques.

C. Semi-supervised Learning

Semi supervised learning is used in many practical learning applications such as text processing, video indexing, and bioinformatics where large supply of unlabeled data is provided but some labeled data is expensive to generate which is limited .So in semi supervised learning, training data is both labeled and unlabeled data. Learning model must learn the structures to organize the data as well as make predictions. When cost associated with labeling is too high to allow for a fully labeled training process then semi-supervised learning is useful. Supervised learning can be used with methods such as classification, regression and prediction. Example is to identify a person's face on a web cam.

D. Reinforcement Learning

In reinforcement learning, learner interacts with a dynamic environment and performs a certain goal without the intervention of a teacher who tells whether it has come close to its goal. With reinforcement learning, leaning technique is used to discover the actions through trial and error and predict which actions yield the greatest rewards. Example is chess playing, learner learns how to play a game against an opponent by performing trial and error actions to win.

Mainly three primary components such as the learner, the environment and actions are used in reinforcement learning. The primary goal of the learner to choose actions that gives the expected result over a given amount of time. So that best policy is chosen by learner to reach the goal in this type of learning.

VI. MACHINE LEARNING ALGORITHMS

Machine learning models are build by implementing large set of algorithms on the basis of learning style and classified as follows:

A. Regression algorithms

Regression is the technique of modeling the relationship between continuously varying variable such as a price, a temperature and refined it iteratively using a measure of error. The most popular regression algorithms are linear regression, logistic regression.

B. Instance-based Algorithm

It is learning model applied on a decision problem which takes instances of training data as input and compare test data using a similarity measure to find the best match and make a prediction. Instance-based methods are called as lazy learner because it simply stores training data and waits until test data is given to perform the learning. So it takes less time in training but more time in predicting. The most popular instance-based algorithms are k-Nearest Neighbour (kNN), Self-Organizing Map (SOM).

C. Decision Tree Algorithm

In this algorithm, decision tree is used as a predictive model which maps observations on input data to predict the item's target value. In these tree structures, leaves are used to represent class labels and branches shows the features that lead to those class labels. Decision trees are fast and accurate algorithms which are trained on data for classification and regression problems. Classification and regression Tree (CART), Chi-squared Automatic Interaction Detection (CHAID) are popular decision tree algorithms.

D. Bayesian Algorithms

Bayesian algorithms are based on probability theory which is used to represent uncertainty. These explicitly apply Bayes' Theorem for problems such as classification and regression. The most popular Bayesian algorithms are Naive Bayes, Gaussian Naive Bayes, Multinomial Naive Bayes, Bayesian Belief Network, Bayesian NetworK.

E. Clustering Algorithm

It is used to classify objects into different groups. Clustering is the type of unsupervised learning where It partitions the data set into clusters which share same characteristics and based on some defined distance measure. Clustering methods are classified as hierarchical clustering and partitional clustering. The most popular clustering algorithms are k-Means, k-Medians, Expectation Maximisation (EM), Hierarchical Clustering

F. Artificial Neural Network Algorithms

This learning algorithm is based on supervised learning and similar to the structure of biological neural networks. Artificial neurons are highly interconnected among units and learning is performed by updating the connection weights to perform parallel distributed processing. The most popular artificial neural network algorithms are Perceptron, Back-Propagation, Hopfield Network, Radial Basis Function Network.

G. Deep Learning Algorithms

Artificial neural networks are updated to give abundant cheap computation to form deep learning. Deep learning algorithms are build on much larger and more complex neural networks to implement semi- supervised technique

where large datasets contain very little labeled data. The most popular deep learning algorithms are Deep Boltzmann Machine (DBM), Deep Belief Networks (DBN), Convolutional Neural Network (CNN).

H. Dimensionality Reduction Algorithms

When object is described with number of dimension, Dimensionality reduction method is used remove irrelevant and redundant data to reduce the computational cost. The dimensionality reduction algorithms are Principal Component Analysis (PCA), Principal Component Regression (PCR), Linear Discriminant Analysis (LDA), Mixture Discriminant Analysis (MDA), Quadratic Discriminant Analysis (QDA), Flexible Discriminant Analysis (FDA).

VII. APPLICATIONS AND TOOLS

Based on learning styles such as supervised and unsupervised Learning, machine learning applications are classified Classification problems like pattern recognition, face recognition, character recognition, medical diagnosis, web advertizing make use of supervised Learning. Applications based on unsupervised learning are clustering, association analysis, customer segmentation in CRM, image compression, bioinformatics. Reinforcement learning applications are robot control and game playing. Selection of right tool is important in machine learning to work with the best algorithms. Good tools are applied in machine learning to make faster, easier predictions. Intuitive interface onto the sub-tasks is provided by machine learning tools by providing good mapping and suitability in the interface for the task. Best practices for process, configuration and implementation are provisioned by great machine learning tools. Automatic configuration of machine learning algorithms is performed and good process is built to structure of the tool. ML tools are divided into platforms and libraries**.** A platform provides environment needed to run a project, whereas a library only provides collection of modeling algorithm needed to complete a project. Examples of machine learning platforms are WEKA Machine Learning Workbench, R Platform, Subset of the Python SciPy like Pandas and scikitlearn. Examples of machine learning libraries are scikit-learn in Python, JSAT in Java, Accord Framework in .NET.

VIII. CONCLUSION

This paper provides a review of machine learning techniques used for early prediction of oral cancer. Also overview of various machine learning techniques and algorithm is given and findings regarding different machine learning approaches applied in the detection of oral cancer are discussed.

REFERENCES

[1] K. Anuradha,K Sankaranarayanan, “Detection of Oral Tumor based on Marker Controlled Watershed Algorithm”, International Journal of Computer Applications (0975 – 8887),Volume 52– No.2, August 2012

[2] Siow-Wee Chang, Sameem Abdul Kareem, Amir Feisal, MericanAljunid Merican, Roshan Binti Zain, “A Hybrid Prognostic Model for Oral Cancer based on Clinicopathologic and Genomic Markers”, Sains Malaysiana 43(4) (2014): 567–573

[3] Wafaa K. Shams , Zaw Z. Htike, “Oral Cancer Prediction Using Gene Expression Profiling and Machine Learning”, International Journal of Applied Engineering Research ISSN 0973-4562 ,Volume 12, Number 15 (2017) pp. 4893-4898

[4] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, Dimitrios I. Fotiadis,

[5] “Machine learning Applications in cancer prognosis and prediction”, Computational and Structural Biotechnology Journal 2015) 8–17

[6] Narayan Naik, Anusha Amin, Dechamma K.C, Deepthi, B.A, Nidhi Hegde, “Oral Cancer Detection Using Android Application”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 1,January 2015.

[7] K. Anuradha, K Sankaranarayanan, “Identification of Suspicious Regions to Detect Oral Cancers at an earlier stage– a literature survey”, International Journal of Advances in Engineering & Technology, March 2012.

[8] Shikha Agrawal, Jitendra Agrawal, “Neural Network Techniques for Cancer Prediction: A Survey, Procedia Computer Science 60(2015) 769-774

[9] Hakan Wieslander, Gustav Forsli, Ewert Bengtsson, “Deep,Convolutional Neural Networks For Detecting Cellular Changes Due To Malignancy”, IEEE Xplore

[10] Neha Sharma ,Hari Om, “Framework for Early Detection and Prevention Of Oral Cancer using Data Mining”, International Journal of Advances in Engineering & Technology, Sept 2012.

[11] Fatihah Mohd, Noor Maizura, Mohamad Noor “Analysis of Oral Cancer Prediction using Features Selection with Machine Learning”, ICIT 2015 The 7th International Conference on Information Technology.