An Efficient Decision making of oral cancer Using fuzzy based machine learning algorithm in Hadoop Framework

[1]M.Ramesh , [2]Dr.C. Dhanalakshmi
[1,2]Prist university, Puducherry Campus

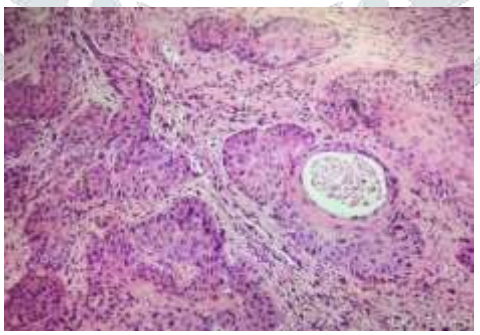
Abstract **:** Oral squamous cell carcinoma (OSCC) constitutes is one of the major type malignant and its significant percentage of oral squamous cell carcinomas and are one of the most common causes of death worldwide. Diagnosis, prediction and control of the OSCC are practically done traditionally based on the clinical signs and historical highlights and biomarkers. Because of absence of convenient diagnosis on the whole ordinary methods or differential diagnosis biopsy of the patient is required. Therefore the patient may be feel inconvenient and lack of knowledge for the hospital staffs or doctors to predict the cancer. In our research, our primary aim to predict and efficient decision making of cancer and its treatment by the historical patient data set and symptoms like smoking and drinking. We use fuzzy based machine learning algorithm to classify and predict the OSCC. The presented system is expected to validly improve the healthcare services by considering the methods.

Index Terms : **OSCC, Diagnosis, fuzzy, biopsy, cancer**.

1. Introduction

Oral Squamous Cell Carcinoma (OSCC) consists of a one of the most incessant neoplasm in people and its death rate is known to be very large. It can be identified in any piece of the oral cavity or oropharynx. It might be alluded to any threat that has been started in the head and neck region . Due to locoregional repeat in malignancy, early recognizable proof of an illness backslide can be pivotal for the patient's forecast and treatment.

One of main factors Oscc is tobacco usage has become more common in India. Smoking, chewing and drinking have been associated with oral lesions such as OSCC and oral submucous fibrosis (OSMF), which has the potential for malignant transformation. According to the study conducted by Nuefold and his coworkers, using national sample survey (NSS) which is representative sample of India, conducted among 4,71,143 people 10 years and older, the prevalence of regular use of alcohol is 4.5%, smoking tobacco is 16.2%, chewing tobacco is 14%.The prevalence of habit was found to be more among men when compared to women, also prevalence was higher in rural population.



Microscopic pathology image showing Squamous Cell Carcinoma of the Oral Cavity.

With the coming of high-throughput advances, for example, DNA microarrays, the articulation levels of numerous qualities can be measured at the same time. The measure of such sample information can be additionally broke down keeping in mind the end goal to recognize considerable arrangements of individual genes with their appearance esteems which give learning in regards to the condition being examined. Besides, by gathering and analyzing , at an functional level, littler arrangements of related qualities enable us to recognize: (i) gatherings of qualities that capacity on the same or potentially unique pathways, and (ii) pathways that vary between two phenotypes .

Evaluation of Salivary Sialic Acid in Patients With Different Clinico-Pathological Stages of Oral Leukoplakia and Oral Squamous Cell Carcinoma - A Cross Sectional Study and concluded that total sialic acid ,lipid bound sialic acid and TSA to total protein ratio plays a vital role in development of oral cancer and leukoplakia.

The most established technique to predict OSCC repeat is the tumor, node and metastasis (TNM) organizing, which is construct predominantly with respect to the dimensional qualities of the tumor and on presence , number, and site of neck nodes metastasis. Tragically, its insufficiency is today, squamous is vary their behavior of uncertain health condition, Which can be now and then extremely forceful and others can metastasize gradually after surgery. This instability in movement has driven researchers to look for a bigger number of markers.

Numerous clinical, histopathological, radiological, and hereditary variables were contemplated, yet none of the diverse gatherings taken, particularly gives clinically material markers of tumor forcefulness. Given the multilevel idea of disease (qualities, cells, tissues, and organs), coordination of the distinctive gatherings of information are required. Though extraordinary reports are available in the writing on information incorporation and formation of institutionalized prognostic calculations for bladder and breast cancer, nothing is accessible for head and neck malignancy.

2. Literature Survey

Konstantina Kourou[1] summarized the study we abuse differentially communicated genes so as to additionally perform pathway improvement analysis . As per our outcomes we discovered huge pathways in which the diseases related genes have been distinguished as firmly improved. Moreover, in light of the after effects of the pathway improvement investigation we propose a technique for predicting oral cancer recurrence utilizing Dynamic Bayesian Networks. The system comprises of three principle steps. In the initial step, transcriptomic information is broke down so as to distinguish a subset of the most differentially communicated genes among the two gatherings of patients. In the second step, pathway enhancement analysis is performed for the particular genes rundown intending to distinguish the most critical pathways as far as overrepresentation. At long last, very associated genes that take part in the most huge pathways are separated.

Dario Salvi..,[2] summarized framework for taking in the structure of oral cancer hereditary system in light of DBNs. Time series gees articulation information are misused to distinguish the most noteworthy qualities and to detail genes interaction networks. Following a genomic analysis in light of computational strategies, DBNs can be utilized for building the causal relationships among the factors inside a similar time-slice and between back to back time-slices We have figured three DBNs: (i) one for the aggregate number of patients, (ii) one for the group of patients that had suffered a relapse and (iii) one for those that had not been determined to have an one for the group of patients that had suffered a relapse. The subsequent quality systems are then dissected as far as topological and useful parameters. The precise prediction of quality communications identified with oral growth can decide conceivable associations of infection states amid various time slices. Additionally, the combination of time-course gene articulation esteems with arrange knowledge permitted us for the revelation of critical nodes identified with oral malignancy recurrence.

Nikolaos Papachristou.., summarized [3] In this paper, a large number of data are produced from the healthcare centers. In advanced analytical and classification technique is called Machine Learning and Data Mining are more popular for providing to complex structure for critical decision making and mechanisms of clinical complex data and are to used construct the solution for decision making for health related data. Although, these scientific leaning methods hace several distinctions from statistics. Machine Learning and Data Mining are more flexible method to use and adopt for both mathematical and heuristic problems on the complex and specific dataset. For example, Based on the computational efficiency and machine learning methods are able to handle entire population of Indian data samples. Furthermore, while statistics handles quantified data, Machine Learning and Data Mining can handle various kinds of data (e.g. CT/MRI images, genomic data, sounds, text, etc.).

Ujjwal Maulik..,[4] summarized this paper, we propose a novel approach that joins fuzzy preference based rough set (FPRS) for discovering more pertinent gene expression data and from microarray quality articulation information and TSVMs to additionally enhance the grouping execution as proposed rough set model is developed in light of comparability relations. The relations have one of the fundamental constraints while applying this model to complex decision tasks. Fuzzy preference relations can mirror the level of inclination quantitatively making it all the more effective in extricating data from fuzzy information than identicalness or predominance relations. This rouses us to utilize FPRS strategy for decision choice. Accordingly, we have connected TSVMs on the subset of quality markers to enhance precision. Utilizing six benchmark quality microarray datasets, the adequacy of the proposed plot is contrasted and SNR and CBFS techniques. Additionally, we have researched diverse mixes of highlight choice techniques and regulated classifiers (SVM and k-NN) utilizing cross-validation exactnesses demonstrate the viability of FPRS in extricating gene markers.

Niharika G. Maity..,[5] paper we examine two such applications for translating medicinal information for robotized analysis . Our first contextual analysis shows the utilization of Bayesian Inference, a worldview of machine learning, for diagnosing Alzheimer's disease in light of cognitive outcomes and statistic information. In the second contextual investigation we concentrate on automated characterization of cell pictures to decide the progression and seriousness of breast growth utilizing artificial neural systems. Despite the fact that these examination are as yet preparatory, they show the estimation of machine learning systems in giving fast, proficient and automated data investigation. Machine learning offers trust with early conclusion of illnesses, help patients in settling on treatment choices on treatment alternatives and can help in enhancing general nature of their lives. The Bayesian Regularization training technique had a more noteworthy than 90% exactness. The Bayesian Regularization technique created a well summing up organizes. Levenberg-Marquardt Optimization strategy was the fastest to prepare. Nonlinear

strategies, other than Bayesian Regularization, just refresh the weights and predisposition esteems yet not a mix of squared mistakes and weights as in Bayesian Regularization technique. Bayesian Regularization technique set aside long opportunity to prepare anyway it is minimum inclined to overtraining.

3. Problem statement

- The most important challenges are that medicine is practiced in a safety critical context in which decision-making activities should be supported by explanations.
- Medical big data can be costly due to involvement of the personnel, use of expensive instrumentation, and the potential discomfort of the patients involved.
- The oral cancer or OSCC causes major problem in human life because there is lack of knowledge about decision of cancers stage ,prescriptions and its technology.
- DNA microarrays, the expression of the genes may be varying their character in the time of biopsy. So the proper processing for required to predict the cancer.

4. Proposed Methodology

The exponential production of data in recent years has introduced a new area in the field of information technology known as 'Big Data'. In a clinical setting such datasets are emerging from large-scale laboratory information system (LIS) data, test utilization data, electronic medical record (EMR), biomedical data, biometrics data, gene expression data, and in other areas in the healthcare. Massive datasets are extremely difficult to analyze and query using traditional mechanisms, especially when the queries themselves are quite complicated. In effect, a Map Reduce algorithm maps both the query and the dataset into constituent parts.

We already discussed about the problem of prediction and decision making about the oral cancer in various algorithm. we propose new novelty algorithm for efficient prediction and decision taking about the symptoms and treatment for the cancer. In our research work the efficient fuzzy based machine learning algorithm are used for proficient classification and analyses.

Fuzzy Logic (FL) is a method of reasoning that resembles human reasoning. The approach of FL imitates the way of decision making in humans that involves all intermediate possibilities between digital values YES and NO.

The conventional logic block that a computer can understand takes precise input and produces a definite output as TRUE or FALSE, which is equivalent to human's YES or NO.

The inventor of fuzzy logic, Lotfi Zadeh, observed that unlike computers, the human decision making includes a range of possibilities between YES and NO, such as

Conditions****
CERTAINLY YES
POSSIBLY YES
CANNOT SAY
POSSIBLY NO
CERTAINLY NO

In case of our research we use fuzzy logic our check the condition of the patient. For example, the following data is based on the stages of the oral cancer.

Evolved from the study of pattern recognition and computational *learning* theory in artificial intelligence, *machine learning* explores the study and construction of *algorithms* that can learn from and make predictions on data Machine learning algorithms can be divided into 3 broad categories

1. Supervised learning,
2. Unsupervised learning, and
3. Reinforcement learning

1. Supervised learning is useful in cases where a property (*label*) is available for a certain dataset (*training set*), but is missing and needs to be predicted for other instances.

2. Unsupervised learning is useful in cases where the challenge is to discover implicit relationships in a given *unlabeled* dataset (items are not pre-assigned).

3. Reinforcement learning falls between these 2 extremes there is some form of feedback available for each predictive step or action, but no precise label or error message. Since this is an intro class, I didn't learn about reinforcement learning, but I hope that 10 algorithms on supervised and unsupervised learning will be enough to keep you interested.

The popular and efficient algorithm for classification of our data is

1. Decision Tree
2. KNN
3. Random Forest

By using any of these above machine learning algorithm we classify our data set efficiently and predict the decision taken for the treatment, medicine based on the symptoms.

4.1 Oral Cancer

i) T stage: The main tumor mass

Based on a physical examination and review of any imaging, your doctor should be able to give you a T stage that falls within one of the following categories.

Tx	The doctor is unable to assess the primary tumor.
T0	The doctor is unable to find the primary tumor.
Tis	Carcinoma in situ (or severe dysplasia); this means there are cancer type cells, but they have not yet invaded deep into tissue. This is more of a pre-cancer lesion.
T1	The tumor is 2 centimeters or less in greatest dimension.
T2	The tumor is more than 2 centimeters but less than or equal to 4 centimeters in greatest dimension.
T3	The tumor is more than 4 centimeters in greatest dimension.
T4a	This is moderately advanced local disease. The tumor clearly invades into the skin of the face, through the upper or lower jawbone, into the nerve that allows you to feel the teeth and chin area or into the floor of the mouth.Note: A little bit of bone or tooth socket invasion from a tumor of the gums does NOT make it a T4a cancer.
T4b	This is very advanced local disease. This stage is assigned if the tumor is invading into the masticator space, pterygoid plates, base of the skull and/or encases the carotid artery.

ii) N stage: Spread of cancer to the lymph nodes in the neck

Next, your doctor will use all the available information and assign you an N stage. This is based on the assessment as to whether the cancer has spread to lymph nodes in the neck.

Nx The neck lymph nodes cannot be assessed.
N0 There is no evidence of any spread to the nodes.
N1 There is a single node, on the same side of the main tumor, that is 3 centimeters or less in greatest size.

N2a Cancer has spread to a single lymph node, on the same side as the main tumor, and it is more than 3 centimeters but less than or equal to 6 centimeters in greatest dimension.

N2b There are multiple lymph nodes that have cancer on the same side as the main tumor, but none are more than 6 centimeters in size.

N2c There are lymph nodes in the neck on either the opposite side as the main cancer or on both sides of the neck, but none are more than 6 centimeters.

N3 There is spread to one or more neck lymph nodes, and the size is greater than 6 centimeters.

iii) M stage: Spread of cancer outside the head and neck

Finally, based on an assessment on the entire body, you will be assigned an M stage.

M0 No evidence of distant (outside the head and neck) spread.

M1 There is evidence of spread outside of the head and neck (i.e., in the lungs, bone, brain, etc.).

This is the stages and case of the oral cancer. By using the fuzzy logic and machine learning to classify the stages based on the attributes in the data set. Finally we predict and will decide the treatment of the patient to save their life.

5. Cancer Stage

Stage 0	Tis	N0	M0
Stage I	T1	N0	M0
Stage II	T2	N0	M0
Stage III	T3	N0	M0
	T1	N1	M0
	T2	N1	M0
	T3	N1	M0
Stage IVA	T4a	N0	M0
	T4a	N1	M0
	T1	N2	M0
	T2	N2	M0
	T3	N2	M0
	T4a	N2	M0
Stage IVB	Any T	N3	M0

	T4b	Any N	M0
Stage IVC	Any T	Any N	M1

5.1 Objectives Of The Research

In India, 72% of the population resides in rural areas and 30-40% of cancers are found in the oral cavity. The majority of residents live in villages where inadequate medical facilities, no proper primary care infrastructure or cancer screening tools and high levels of illiteracy all contribute to poor oral cancer (OC) outcomes. In this challenging environment, the objective of this study was to assess the association between various risk factors for OC among referrals for suspicious lesions. Oral cancer is of significant public health importance to India. Firstly, it is diagnosed at later stages which result in low treatment outcomes and considerable costs to the patients who typically cannot afford this type of treatment. So, our research based on the help to health services in the India. We study the different prediction methods and finally solve the efficient results by proficient algorithms This method help to health sectors by predict the cancer stages and will give good treatment to the patients in the efficient manner.

Results

Site	**Oral Cavity**
Subsite	**Tongue (Right Lateral Border)**
Type	**Squamous Cell Carcinoma**
cT	**cT2**
cN	**cN0**
cM	**cM0**
cStage	**cII**

6. Conclusion

In conclusion, we presented the new methodology for predicting and decision making of oral cancer or OSCC using fuzzy based machine learning algorithms. In India large number of patients suffered from the oral cancer by the consuming alcohol, smoking etc. As a research scholar, we decided to help to the health service by our proposed methodology. During the steps of the proposed system, the fuzzy logic is used to catagories the patients by the different parameters, based on the results machine learning algorithms is used to classify the dataset and predict the results for decision making to the patients

REFERENCES

[1] Konstantina Kourou, Costas Papaloukas and Dimitrios I. Fotiadis, Integration of Pathway Knowledge and Dynamic Bayesian Networks for the Prediction of Oral Cancer Recurrence, IEEE Journal of Biomedical and Health Informatics,IEEE journal of Biomedical and Health Informatics, Volume: 21, Issue: 2, Pages: 320 – 327,Year:2017.

[2] Dario Salvi∗, Marco Picone, Mar´ıa Teresa Arredondo, Mar´ıa Fernanda Cabrera-Umpi´errez, Merging Person-Specific Bio-Markers for Predicting Oral Cancer Recurrence Through an Ontology, Ieee Transactions On Biomedical Engineering, Vol. 60, No. 1, January 2013.

[3] Nikolaos Papachristou, Christine Miaskowski, Payam Barnaghi, Roma Maguire, Nazli Farajidavar, Bruce Cooper and Xiao Hu, Comparing Machine Learning Clustering with Latent Class Analysis on Cancer Symptoms' Data, 2016 IEEE Healthcare Innovation Point-Of-Care Technologies Conference (HI-POCT) , Pages: 162 – 166, Year: 2016.

[4] Ujjwal Maulik, and Debasis Chakraborty, Fuzzy Preference Based Feature Selection and Semisupervised SVM for Cancer Classification, Ieee Transactions On Nanobioscience, Vol. 13, No. 2, June 2014.

[5] Niharika G. Maity; Sreerupa Das ,Machine Learning For Improved Diagnosis And Prognosis In Healthcare ,2017 IEEE Aerospace Conference ,Pages: 1 - 9, Year: 2017.