May - June 2020
ISSN: 0193-4120 Page No. 4474-4477

Prediction of Oral Cancer using Data Mining

Rachana C[1], Pooja Gorantla[2], Snehal Rathi[3], Lavanya J S[4], Manju More E[5]

[1,2,3,4]UG Student, [5]Professor,
[1,2,3,4,5]School of Computing & IT, Reva University, Bangalore, Karnataka
[1]rachanareddyc7@gmail.com, [2]poojagorantla2000@gmail.com, [3]snehalrathi112000@gmail.com,
[4]lavanyapsv@gmail.com, [5]manjumore.e@reva.edu.in

Article Info
Volume 83
Page Number: 4474-4477
Publication Issue:
May - June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *12 May 2020*

Abstract
Cancer is one of the dangerous disease. Its growth rate is very high. Among all the types of cancer oral cancer is can be considered as one of dangerous cancers which originates from the oral cavity& neck. Overuse of tobacco and smoking cigarettes is the primary reason for developing oral cancer. India has the highest number of oral cancer patient. In this project we are using data mining technology. Some of the data mining technology are classification algorithm, clustering method etc. Among all these technology a suitable method is choosen for better and easy classification of data. These techniques gives relationship between the data.

Keywords: *Oral Cancer, Deep Neural Network, Adaptive Fuzzy System, C-Means Clustering*

1. Introduction

Death rate of oral cancer is high(2,00,000 annually over the world &46,000 annually in India). In a research it is found t that the developing countries have the high rate of oral cancer than the developed countries. Good dental oral care is important to maintain healthy teeth gums and tongue. Moral problem including bad breath , TMD, tooth decay, or thrush are all the treatable with proper treatment and care. Oral cancer can affect any area including the lips gum tissue tongue cheek teeth. This leads to the most deaths globally. It is possible to easily reduce deaths caused by cancer, by detecting it early stages and give the proper treatment.

The approach used here is described by analyzation. The dataset consists of patient's habits, symptoms and medical history. It is possible to discover patterns in large sets of data through data mining which is a computational process. Data mining uses a combination of techniques related to database systems, statistics, Artificial Intelligence & machine learning.

Data mining helps to extract and transform data in a dataset. Data Mining acts as an important role in the Prediction of Cancer Diseases. Data mining methods are mainly used to develop a predicted model in medicine field. By testing several data mining methods, we have to choose the model which gives high accurate results. In the classification model the main metrics are done using the Sensitivity, Specificity, Accurateness, Error Rate, True Positive Rate (TPR) and False Positive Rate (FPR).Oral cancer is located in oral cavity. It can originate in any of the oral tissues or by extension from a neighboring architectural structure that is the nasal cavity.

2. Literature Survey

Kent (2007) used the method of genetic programming to solve complex problems. This technique is used to collect sample of patients and develop program accordingly to detect oral cancer. The sample of data of patients consist of details of their habits, lifestyles and medical history. Although there are less amount of positive samples in this method it is able to provide more accurate results. This can compete with previous technology.

- Nahar (2011) discussed the measures to prevent from the particular type of cancer. To find out the steps factors they have first collected dataset. They used three association algorithms: 1.Apriori 2.Predictive Apriori and 3.Tertius algorithms. These algorithms are helpful to discover the important factors against a specific type of cancer. From the analysis it is found that Apriori algorithm is the most efficient association algorithm to take measures for prevention.

- Chuang et al. (2014) used the method of DNA repair genes. They used single nucleotide polymorphisms (SNPs) data set of oral cancer patients (238 samples) for analyses purpose. The support vector machine was used to conduct experiments and they analysed that the

performance of the holdout cross validation (test and train) were better than tenfold cross validation (rotation estimation). The accuracy was upto 64.2 %.The people named Gadewal and Zingde took this method forward by adding 132 genes to 238 samples making it to 374 gene database and tried to enable fast retrieval of updated information.

- Kaladhar et al. (2011) used the method of classification algorithm (CART, Random Forest, LMT and Na¨ıve Bayesian algorithms).In these algorithms classification is done on the basis of rotation & estimation and training dataset. Among these algorithms Random Forest classification technique is more efficient.
- Sankaranarayanan (1995) studied the aspects of the oral cancer in India. The causes such as chewing of tobacco or any tobacco stem or leave are studied in detail. According to author study the age frequency is a decade earlier than the age mentioned in western literature. Only upto 10-15% is limited from where it is originated.
- Anuradha & Sankaranarayanan (2012) have carried out survey on many major methods adopted by the researchers to detect oral cancer at initial stage itself.

Author compared all the methods to determine which method provides more accurate results.

- Milovic (2015) used patterns to detect cancer at initial stage so that physicians can treat the patients at earlier stage itself. The pattern recognization is also a data mining technology.
- Gadewal and Zingde (2013) enhanced the oral cancer gene database to include 374 genes by adding 132 gene entries to the 238 samples given by chuang to enable fast retrieval of updated information.
- Gupta et al.(2012) used artificial neural network(ANN). The accuracy of ANN is 93.68% to prepare dataset and 55.5% to accept dataset.

3. Objectives

- The main aim of the current procures is towards developing a new predict the stages of tuner growth in oral cancer. The risk factors of oral cancer consist of certain actions such as tabacoo chewing & alcohol drinking which are considered as the major risk of oral cancer.
- Summer of oral cancer are easily detected if we use the data mining technology it can be predicted at early stages the treatment is success only if it's detected & diagnosed at early stages.
- Unlike breast & colorectal cancer no genetic & measures made exist for oral cancer to predict outcome. The present technology tracks accuracy data mining to build such a model to calculate recurrence by taking into account of various patient & tumor characteristics.

4. Methodology

Discovering patterns computationally in large and complex data sets is referred to as data mining. This method is a combination of machine learning, database systems, artificial intelligence, and statistics. The objective of the process of data mining is the extraction of information from datasets and applying transformations to obtain a structure that is meaningful in future. Large medical data sets are the best candidates for this classification method.

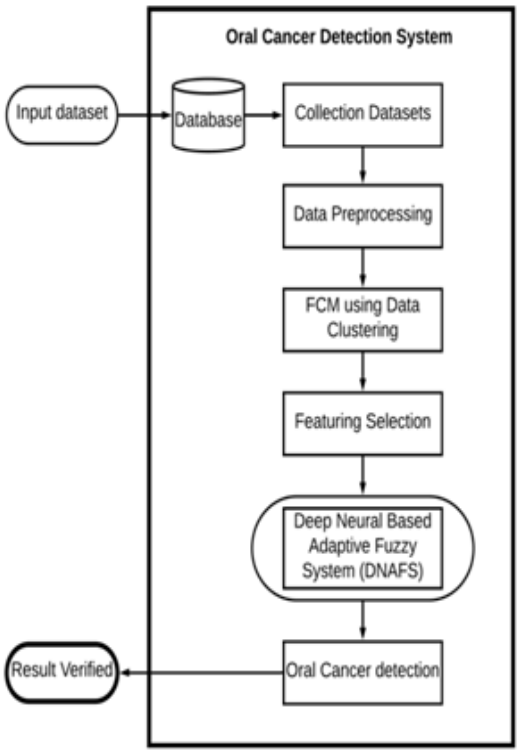

Figure 1: Overall Proposed Architecture

Several data mining techniques are implemented in unison to diagnose and prognoses oral cancer for a specific patient. The exploration of DNAFS and data mining methods is done for the identification of appropriate techniques and methods that efficiently classify data. In the end, classification is done using Deep Neural Network Based Adaptive Fuzzy System (DNAFS) which is a machine learning technique. The prognosis of the affected patient determines whether successful treatment for the diseases is possible. In this context, information for prognosis is generated with the help of a statement of prognosis.

Dataset

There are two sub datasets for this dataset. The first data set has samples of healthy patients and other samples are of cancer patients who were diagnosed of cancer of the oral cavity. This division helps obtain an indication.

May - June 2020
ISSN: 0193-4120 Page No. 4474-4477

It contains the standard index of oral information for tolerant points of interest for oral cancer. Standard highlights are incorporated into this information. The oral cancer information is collected from a specific clinic or different disease organizations. A range of data sets are involved here such as cerebrum, dental, mouth, and neck.

Preprocessing

Data mining techniques are used in the preprocessing stage to analyze a large data set and identify target data. The preprocessing stage consists of a number of tasks such as cleaning of data, its integration, transformation, reduction and discretization. In the data cleaning stage, noise is eliminated and data is made consistent and coherent.

Absent values are included in this process and outliers are also found.

- Data Integration: This is performed with the help of data cubes or files, and many different databases.
- Data Transformation: This refers to normalizing and aggregating data.
- Data Reduction: This refers to a reduction in volume but the production of similar results of analysis
- Data Discretization: This refers to reduction of data and replacement of its numerical attributes with nominal attributes.

Data in the real world is incomplete, inconsistent, and noisy. Preprocessing of data is a regular attempt for filling values which are missing and smoothing out noise. In this stage, outliers are identified and inconsistencies in data are corrected. The missing values are filled manually or through the use of a global constant.

Fuzzy C-Means Clustering

Clustering refers to a task when objects are grouped so that similar objects belong to the same group or cluster. This is the primary task when it comes to exploratory data mining. It is a common analysis technique for statistical data and use in many fields such as pattern recognition, bioinformatics, image analysis, information retrieval, machine learning, computer graphics and data compression. One of the common descriptive tasks is clustering, where a finite set of clusters are identified for data description. Grouping elements with similar characteristics is part of the clustering process. The mean of each cluster is used in the technique. Similar data values are grouped in one cluster.

Fuzzy C- Means clustering technique.

- FCM or Fuzzy C-Means clustering is also referred to as Fuzzy for the medical data.
- Fuzzy Partitioning is part of FCM. In this case, specific data can be part of all groups having different grades of membership between 0and1.
- The nature of FCM is iterative. Its objective is to find centroid or cluster centres so as to minimize the function of dissimilarity.

Algorithm

The algorithm assigns membership to every data point pertaining to a cluster centre. This assignment is based on the distance between the data find point and the cluster centre. If there is more data in the proximity of the cluster centre, then it means that its membership towards that cluster centre is greater.

1. Initialize M=[s_{ij}] matrix, $M^{(0)}$
2. At k-step: calculate the centers vectors $C^{(k)}=[c_j]$ with $M^{(k)}$

$$C_j = \frac{\sum_{i=1}^{n} s_{ij}^m y_i}{\sum_{i=1}^{n} s_{ij}^m}$$

3. Update $M^{(k)}$,$M^{(k+1)}$
4. $d_{ij} = \|y_i - c_j\|$, $d_{kj} = \|y_k - c_j\|$,

$$s_{ij} = \frac{1}{\sum_{k=1}^{c} (\frac{d_{ij}}{d_{kj}})^{2/(m-1)}}$$

5. If $\|M^{(k+1)}-M^{(k)}\| < \epsilon$ then Stop;

Else

Return to step 2 ;

Here M is any real number greater than 1,
s_{ij} is the degree of membership of y_i in the cluster j,
y_i is the ith of d-dimensional measured data,
c_j is the d-dimension center of the cluster.

Figure 2: Algorithm

5. Result

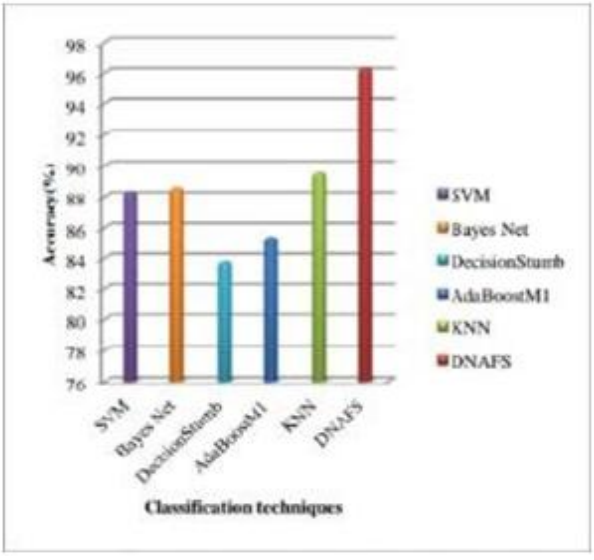

Figure 3: Performance of accuracy result

6. Conclusion

In healthcare, the concept of data mining is becoming more essential. In this paper, various methods are studied to detect cancer .The research work applies data mining

May - June 2020
ISSN: 0193-4120 Page No. 4474-4477

for oral cancer risk and divides in terms of age, gender and socioeconomic status. The main aim of this model is to provide more accurate results to the users and it is also cost and time saving to the user. The age group which is suspectible to oral cancer is middle age because of change in lifestyle. This technology will help to detect oral cancer at initial stage so that it will be easy for doctors to give proper treatment at initial stage itself.

7. Future Work

In the future Innovation tumours will play a important role for better treatment methods. Data mining techniques can be used to identify the stages and treatments of oral cancer. By using present technology better tool can be discovered for easy detection and more accurate result of disease. In future neural network can be build using present technology.

Reference

[1] Charles Edeki "Comparative Study of Data Mining and Statistical Learning Techniques for Prediction of Cancer Survivability" Mediterranean journal of Social SciencesVol 3 (14) November 2017.

[2] Vikas Chaurasia "Data Mining Techniques: To Predict and Resolve Oral Cancer " International journal of Computer Science and Mobile Computing (IJCSMC), Issue. 1, January2018.

[3] Reeti Yadav "Chemotheraphy Prediction of Cancer Patient by Using Data Mining Techniques" International Journal of Computer Applications (0975-8887), Volume 76-No.10, August2018

[4] Nikhil Sureshkumar Gadewal, Surekha Mahesh Zingde, "Database and interaction network of genes involved in oral cancer", (2019).

[5] Jaya Suji. R, Dr. Rajagopalan S.P, "An automatic Oral Cancer Classification using Data Mining Techniques", International Journal of Advanced Research in Computer and Communication Engineering, in2019.

[6] S. Warnakulasuriya, Global epidemiology of oral and oropharyngeal cancer", April-May2009.

[7] I. Van Der Waal and R. De Bree, "Second primary tumours in oral cancer", June2010

[8] Tasnuba Jesminet.al. "Brain Cancer Risk Predication tool using data minig", January2013.

[9] K Anuradha, Dr K Sankaranarayanan "International Journal of Advance Research in Computer Science and Software Engineering" issue 1 January2015.

[10] Sharma, N. And Om, H.(2016) Framework for early Detection and Prevention of Oral cancer using Data Mining International Journal of Advances in Engineering and Technology, 4, 302-310.