ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

An automatic Oral Cancer Classification using Data Mining Techniques

Jaya Suji. R[1], Dr.Rajagopalan S.P[2]

Master of Computer Applications, Sathyabama University, Chennai, India [1]

Professor, Master of Computer Applications, MGR University, Chennai, India[2]

Abstract: In recent days, India has the uncertain distinction of harbouring the worlds most number of oral cancer patients with and annual age standardized occurrence of 12.5 per 100,000. Normally, the head and neck cancer (H & N cancer) is the 6th most common cancer in all over the world. Oral cancer is generally analysed type of head and neck cancer is more and more globally in incidence and developing seriously in various region of the world. In the proposed work of this research, the datasets are obtained from different diagnostic centers which contains both cancer and non-cancer patients information and collected data is pre-processed for duplicate and missing information and also to proposes a three various classification algorithm are utilized that is C 4.5, Random tree, and multilayer perceptron neural network model. The best accuracy for the given datasets is achieved in C4.5 algorithm match up to other Classification algorithm and also predictions of oral cancer. Also it is separate to identify the cancer and non-cancer patient's data set records. The data mining methods and techniques will be explored to identify the suitable methods and techniques for efficient classification of NMDS dataset. Finally, in this proposed scheme assists doctors or specialist, researchers in their diagnosis decisions and also in their treatment planning process for various categories.

Keywords: Oral cancer, C4.5, MPNN, Random tree, Classification, Data mining, Prediction

I. INTRODUCTION

Oral cancer is a usually identified type of head and neck cancer, which is rising globally in occurrence and growing critically in many region of the world. In 2009, the oral cancer is declared as the 6th most usual cancer in all over the world [1, 2]. In 2008 worldwide, an estimated 263,900 novel cases and 128,000 deaths are happened [3]. The maximum rates are found in Eastern Europe, Melanesia, South-Central Asia and Central, at the same time as the minimum rates are in Eastern Asia, Central America, and Africa for both females and males [3]. Normally, the most affected of oral cancer in some countries like as Bangladesh, Pakistan, India, and Sri Lanka and it will affect up to 25 percent of all new cancer cases [1]. Oral cancers are diagnosed by based on some symptoms that are discussed in below.

A.Oral Cavity and Oropharyngeal Cancers Diagnosis

Pre-cancers or some cancers of the throat and mouth will be found during a test by a dentist or a doctor. But a lot of these types of cancers are found for that reason of symptoms or signs of a patient are having. Suppose if cancer is suspected, tests may be required to confirm the diagnosis.

Symptoms and signs of oral cavity or oropharyngeal cancer [4] [5]

Possible Symptoms and signs of these cancers can consist of:

- The most common symptom that is painful in the mouth that doesn't cure.
- Also very common symptom that is pain in the mouth that doesn't move
- A thickening or lump in the cheek
- A red or white patch on the tonsil or lining, gums, tongue, of the mouth.
- A feeling or a painful throat that something is caught in the throat, but it does not move.
- Problem in swallowing or chewing
- Problem in tongue or moving the jaw
- Other area of the mouth or Numbness of the tongue
- Become uncomfortable or swelling of the jaw that causes dentures to fit poorly.
- Jaw or Loosening of the teeth or pain around the teeth
- Voice changes
- Weight loss
- Mass in the neck or a lump
- Constant bad breath.

Several of these symptoms and signs can also be affected by benign problem, less serious, otherwise even by other cancers. Also still, it is very significant to see a dentist or doctor, suppose if any of these conditions lasts more than

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

two weeks in order that the affect can be found and treated, if required.

However, the oral cancer is to predict cancer and non-cancer of the each every patient by using various classification techniques. In this research of the proposed work, the datasets are get from various diagnostic centres contains both cancer and non-cancer patient's information and collected data is pre-processes for duplicate and missing information and inconsistent of data and also to propose a three different classification algorithm are utilized. [i.e. C 4.5, Random tree, and MPNN (Multi-layer Perceptron Neural network) model]. The best accuracy for the given datasets is achieved in C4.5 algorithm match up to other Classification algorithm and also predictions of oral cancer. Also it is separate to identify the cancer and non-cancer patient's data set records. Finally, in the proposed approach helps doctors in their diagnosis decisions and also in their treatment planning procedures for different categories. This paper come across various steps which are of i) The related works ii) The proposed work about the different classification algorithm iii) Experimental Results are been discussed iv) the conclusion and the further research are been discussed.

II. RELATED WORKS

A number of works has been focused on different diseases like cancer that is explained in the paper [6]. The suitable techniques are used since the Decision Tree is easy to recognize, handles classification, models non-linear functions, works with mixed data types and the appropriate tools used in [7]. In [8] Dr. Y.S. Kumaraswamy and Shantakumar B. Patil used the MAFIA and K-means clustering algorithm on Heart Disease in order to mine the hear disease that is occurred frequently. The important frequent patterns weightage are considered. Additionally, significant patterns to heart attack prediction are selected based on the computed weightage. These patterns are used for the purpose prediction of heart attack system. An idea regarding serious diseases and their diagnosis using the technique of data mining with smallest number of attributes and provides the consciousness about the diseases that leads to serious concern in [9] Sudha et al . In [10] K.Balachandran et al identify the cancer diseases that are dangerous in treating and diagnosing the users. Thus it is important that the one who have risk factors and more symptoms are best to undertake the medical that is provided by the specialist.

SEER dataset are used and comparison is made in their study in [11] Delen et.al. it reported about the decision tree algorithm that have a good performance than the other two algorithms such as Logistic regression model and Artificial Neural Network. Logistic model doesn't provide the better accuracy. Also the study has conducted on the datasets of death and survival of cancer patients. The decision tree shows the accuracy. [12] The research paper of Arihito Endo et al presents the suitable models to predict the survival cancer patients in the last five years. Also this study provides the highest accuracy based on the logistic regression model. The ANN consists of highest specificity and J48 consist of highest sensitivity. The decision tree representation shows the maximum sensitivity. The Bayesian representation was appropriate to demonstrate the accuracy. Also they found that the optimal algorithm may be dissimilar by the predicted datasets and objects

III. PROPOSED SYSTEM DESIGN

As a proposed work, this research builds the NMDS datasets are obtained from different diagnostic centre which contains both cancer and non-cancer patients information and collected data is pre-processed for duplicate and missing information and also to proposes a three various classification algorithm are utilized that is C 4.5, Random tree, and multilayer perceptron neural network model. And then separate the cancer and non-cancer of each and every patient by using classification techniques. Finally, obtain the best result.

A. Datasets

The BAHNO **NMDS (National Minimum Data Set)** is to be the smallest amount of data that all expert oncologist or other specialist, doctors, surgeons are managing the patients with (H & N cancer) must be expected to collect the data of each and every patients suffering with (H & N cancer). The standard NMDS should contain the minimum amount of data is required to:

- Describe the patient route for each cancer stage
- Find an accurate patient sub-groups
- Explain each cancer organization
- Capture significant result
- Create survival information.

In the dataset contains number of variables included all the fields depends on the standard medical record type. Here the dataset were prepared totally 23 variables [Table 1] (21 input variables and 2 output variables). There is two numerical variable i.e. Case id and Age and as a Categorical variable, we used Gender (Male, Female), History of Addiction (Alcohol, Smoking, Gutka, None, All), Co-Morbid Condition (Hypertension, Diabetes, Immuno-compromised, None) Symptoms (No, Burning, Ulcer, Mass, Loosening of tooth), Site (BM, LA, RMT, LIP, Tongue, UA, Palate), Gross Examination (Ulcero-proliferative, Infiltrative, Verrucous, Plaque Like, Polypoidal), Predisposing Factor (Leukoplakia, Submucous Fibrosis), Tumor Size (<2cm, 2 cm to 4 cm, >4 cm), Histopathology (Variant of SSC-Verrucous, Papillary, Basaloid, Plaque Like, Sarcomatoid,

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

acantholytic, Lymphoepithelioma like), Neck Node (Present, Absent), LFT (Normal, Deranged), USG (Yes, No), FNAC of Neck Node (Yes, No), Diagnostic Biopsy (Squamous Cell Carcinoma, Variant of SCC, Benign), CT Scan / MRI (Bony Involvement, Normal) Diagnosis (SCC, Verrucous, Benign, Plaque Like, Sarcomatoid, Acantholytic, Adenoca, Lymphoepithelioma Like), Staging (I, II, III, IV), Surgery (Y,N), Radiotherapy (Y, N), Chemotherapy (Y, N).

TABLE 1
Independent Variables and its Categories

No	VARIABLE	Category
1	Patient ID	Numerical
2	Age	Numerical
3	Gender	Categorical
4	History of Addiction	Categorical
5	Co-Morbid Condition	Categorical
6	Symptoms	Categorical
7	Site	Categorical
8	Gross Examination	Categorical
9	Predisposing Factor	Categorical
10	Histopathology	Categorical
11	Neck Node	Categorical
12	LFT	Categorical
13	Tumor Size	Categorical
14	Cancer	Categorical
15	USG	Categorical
16	FNAC of Neck Node	Categorical
17	Diagnostic Biopsy	Categorical
18	CT Scan /MRI	Categorical
19	Diagnosis	Categorical
20	Staging	Categorical
21	Surgery	Categorical
22	Radiotherapy	Categorical
23	Chemotherapy	Categorical

B. *Preprocessing Datasets*

Data Cleaning: The real world data have inconsistent, noisy and incomplete. This noisy information's are identified and cleaned while finding the outliers, correct inconsistencies in the data and missing values.

- *Noisy Data*

Noise is nothing but the variance or the random error occurred in the data. There are many technique was used to eliminate the noise and smooth out the data. The outliers are identified by the clustering techniques, whereas the same values are clustered into single cluster or region, values that are lie outside the set of clusters are measured as outliers. The computer and human inspection approaches are combined together using the clustering approaches and create the group of data sets, the human will classifies the patterns in the list in order to recognize the authenticate the garbage ones. It is too faster than the manually search throughout the whole database.

TABLE 2
Outliers Representation

Tumor Size (in cm)	Neck Nodes	LFT	FNAC of Neck Node	Diagnostic Biopsy	USG	CT Scan/MRI	Diagnosis
>4	Abs	Normal	No	Benign	No	Normal	Benign
<2	Abs	Normal	No	Benign	No	Normal	Benign
>4	Abs	Normal	No	Benign	No	Normal	Verrucous
<2	Abs	Normal	No	Benign	No	Normal	Benign
>4	Abs	Normal	No	Benign	No	Normal	Benign

- *Missing Values*

The missing values are found by the different approach based on the significance of the missing values and its relation to the search domain. Either use the global constant to fill in the missing value or else fill the missing values manually.

TABLE 3
Missing Values Representation

Diagnosis	Staging	Surgery	Radiotherapy	Chemotherapy	Histopatholog
Verrucous	II	Y			Verrucous
SCC	IV	Y	Y	Y	SCC
Benign	No	Y	N	N	
SCC	IV	Y	Y	Y	SCC
Benign	No	Y	N	N	Schwanoma

- *Inconsistent Data*

In the certain transaction, there may be the inconsistencies in the data record. Some of the data inconsistency may be proper manually by means of external references. There also exists some inconsistency because of the data integration, whereas the same data values may be represented by various names or else the specified attribute might have various names in the various databases

TABLE 4
Inconsistency of Data (Tobacco-Smoking and Smoking represent the same value)

Clinical Symptom	History of Addiction	Co Morbid Condition	Gross Examination
Burning Sensation	Tobacco-smoking	Alcohol	Plaque-Like
Mass	None	None	Polypoidal
Burning Sensation	Smoking	Alcohol	Plaque-Like
Burning Sensation	Smoking	Alcohol	Plaque-Like

C. ***Selection of Data Mining Algorithms***

A DM (Data Mining) algorithm is a set of assisting to discover and calculations, which create a data mining type from data. Here we are select the efficient algorithm to

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

utilize for a specific logical task is challenge. When we utilize various algorithms to execute the same task or job, each algorithm creates a various result, and a number of algorithms can produce multiple type of result. So that, we can utilize Decision tree algorithms are applied to the mining of very large real-world databases. It is not only for prediction, but also as a way to rectify the number of columns in a dataset.

- *C 4.5*

The Oral cancer datasets are classified through the C 4.5 algorithms that are fundamentally a suitable group of algorithms. These group algorithms are used in categorization in the purpose of data mining and machine learning. The attribute values of the categories are mapped with help of this C4.5 algorithm. The categories will be utilized for novel and unseen instances i.e. new Oral cancer records. The every row indicates a new Oral cancer record that is described through the attributes. The Iterative Dichotomizer 3 algorithm described the decision tree is a method for associated regular questioning of the attribute and their branches describe the value. The decision tree nodes aim to associate the values for prediction as category variables. Similarly, the C4.5 algorithm constructs a decision trees using a set of training data with help of the concept information entropy. Consider a sample set $S=s_1$, $s_2 \ldots s_n$ and every sample s_i is A_{t1}, $At_2 \ldots At_m$ where ai is the feature/attribute. The C4.5 extracts the top possible attribute to divide the node into category/other.

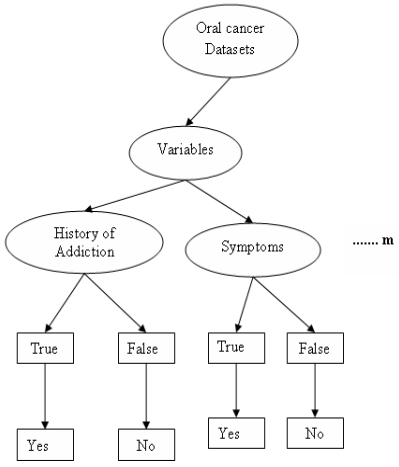

Fig.1 Classical example of C4.5

Totally two branches are constructed by Boolean attribute. If the feature/attribute is categorical the check is having more valued but various values that will be suited into a minimum set of choices with one category predicted for every choice. Previously, categorized data created commonly called samples. If the samples are nominal then the binary values can comes once more into the picture in such a case.

In the Fig.1 we will get to know how the classification is done in the phase 1, where the categorical Oral cancer datasets defines variables of each patient. The variable category consist of defined for History of Addiction, symptoms, etc. next the Boolean variable that defines whether the variables is true or not (i.e the patients affected by the variable or not).

Pseudo code of Decision Trees

1) Test for base cases
2) for every attribute At
3) Discover the normalized data gain from dividing on attribute At
4) Let $At_{\_best}$ be the attribute with the maximum normalized data gain
5) Construct a decision node that divides on $At_{\_best}$
6) Recurse on the sublists acquired by dividing on $At_{\_best}$, and attach those nodes as children

Algorithm: C 4.5

Input: An attribute At which is having valued dataset DS

1) Tree = {}
2) If DS is "empty" OR other terminate criteria met then
3) Stop
4) end if
5) for every attribute At∈DS do
6) Calculate data-theoretic criteria if we divide on attribute At
7) end for
8) At_{best}= Top attribute depend upon above computed criteria
9) Tree = Construct a decision node that checks At_{best} in the root
10) DS_v= Persuaded sub-datasets from DS using At_{best}
11) for every DS_v do
12) $Tree_v$= C 4.5(DS_v)
13) Attach Treev to the equivalent branch of Tree
14) end for
15) Return Tree

- *Random Tree*

Adele cutler and Leo Breiman developed the random tree that provides an extra layer of uncertainty to bagging (both by regression and classification). This tree is used for data item clustering in a set of functionality since it is a statistical algorithm. It is a method to data analysis and predictive modelling. Random tree facilitates outliers and anomaly avoidance and error detection. It achieves best accuracy. The high dimensional data can be easily visualized with the help of random tree. The group of data in tree like arrangement is called as random forest. Random forest consists of exact

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

rules for tree combination, post processing, tree growing and self testing. Random forest algorithm is defined as: The algorithm gets the first randomization via bootstrap aggregation then corresponding bootstrap samples are taken. Thus new training sets are created through random sampling for N'<=N times with replacement.

Next, parallel grouping of learners are separately trained on unique bootstrap samples which is called as Bootstrap aggregation. Final step is classification or mean prediction.
N=training vases; p=variables.

1. From the original data, collect n tree bootstrap samples.

2. An un-pruned regression/classification tree is grown for each of the bootstrap samples with the following adjustment: Instead selecting the best split from all predictors, at each node randomly sample the predictors and select the best split from those variables.

3. New data can be predicted by collecting the predictions of trees.

The algorithm for the Random tree classification algorithm is given below:

Start
{
RF= {Choose attributes subset of given dataset D}
For each chosen variable
{
If (RF.av == True) then take the relevant attributes
Else
Take the unrelevant attributes
}
for all RF until leaf node is reached.
End

Relevant attributes –cancer, Non-relevant- non-cancer

- *Multilayer Perceptron Neural Network Model*

The most commonly used prediction model is Artificial Neural Network (ANN) which is related to human cognitive structure. To solve non-linear problems, different types of ANNs such as Kohonen's self-organizing map, Radial Basis Function Neural Network and multi-layer perception are introduced. These networks solve the problems by learning. The terms Artificial Neural Network (ANN) and Neural Network (NN) generally represents a Multilayer Perceptron Network when used without qualification. A perceptron network consists of three layers is illustrated in the following Fig.2.

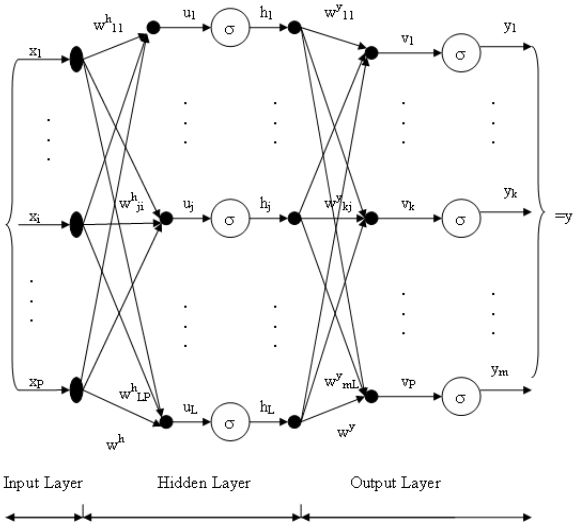

Fig. 2 Multilayer Perceptron Neural Network

The input layer, hidden layer and output layer of this network, each of them consists of three neurons. In the input layer, for each predictor variable, a neuron is assigned. But for categorical variables, N-1 neurons are enough to represent N categorical variables. The above shown network diagram is an example of fully connected, perceptron and feed forward neural network with three layers. Here the term fully connected represents that, the outcome of each input and hidden layer neuron is connected to all of the neurons in the subsequent layer. Feed forward means the input and outcome values only shift from input to hidden and then to output layers; no values are moved from output to input layer (i.e. no back propagation). If the network has more than one hidden layer then the outcome from one hidden layer is given to next hidden layer and unique weights are applied to the sum fed into each layer. Here a multi-layer perceptron neural network was used for organizing the model and the network was split into input, hidden and output layers. The number of neurons in the input, hidden and output layers are 15,2, 2 respectively.

IV. PERFORMANCE EVALUATION AND EXPERIMENTAL RESULTS

In order to evaluate the performance of the algorithm, we carried out the datasets from oral datasets records. Using training data to derive a classifier or predictor and then to estimate the accuracy of the resulting learned model can result in misleading overoptimistic estimated due to overspecialization of the learning algorithm data. Instead, accuracy is better measured on a test set consisting of class-labelled tuples that were not used to train the model. The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. Some of the performance measures are given below. A

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

confusion matrix is a useful tool for analysing classifier accuracy. Structure of confusion matrix is given below.

TABLE 5

Confusion Matrix

		C1	C2
ACTUAL CLASS	C1	True Positives	False Negatives
	C2	False Positives	True Negatives

True Positive (TP) refers to positive tuples that were correctly labeled by the classifier. True Negative (TN) refers to negatives tuples that were correctly labeled by the classifier. False Positive (FP) refers to negatives tuples that were incorrectly labeled by the classifier. False Negative (FN) refers to positive tuples that were incorrectly labeled by the classifier. For an instance, the datasets have trained a classifier to classify medical data tuples as either "cancer" or "non-cancer."

TP – "cancer" tuples that were correctly classified as such

TN – "non-cancer" tuples that were correctly classified as such

FN –"cancer" such as, incorrectly predicting that a cancerous patient is not cancerous

FP – "non-cancer" incorrectly yet conservatively labeling a non-cancerous patient as cancerous

Accuracy: Accuracy is the percentage of tuples that are correctly classified by the classifier.

Accuracy = (TP+TN)/ (TP +TN+FP+FN)

An efficient analysis of classification technique is utilized and it has investigated three data mining techniques: The multilayer perceptron neural network, Random Tree, and the C4.5 decision tree algorithms. Finally, they concluded that C4.5 algorithm has a much better performance than other two techniques and the performance graph is shown in below Fig.3.

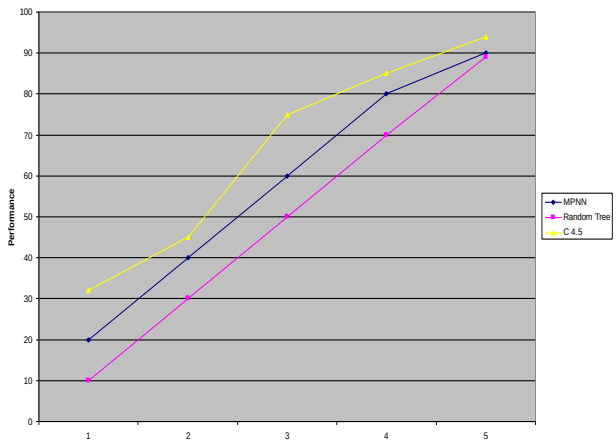

Fig. 3 Performances of Classification algorithms

Fig. 4 shows the oral cancer detection of both cancerous and un-cancerous patient's datasets are resulted by C 4.5. Compared to other classification algorithm C 4.5 obtains best results.

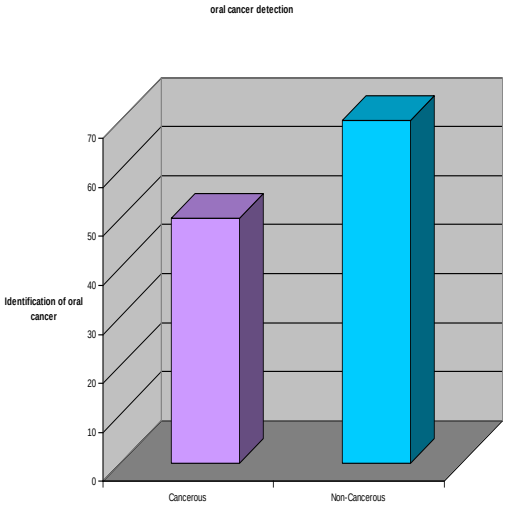

Fig. 4 Prediction of oral cancer

TABLE 6
Accuracy of Classification Algorithms

CLASSIFICATION ALGORITHMS	ACCURACY (%)
C 4.5	100
MPNN	99.5
RANDOM TREE	98.7

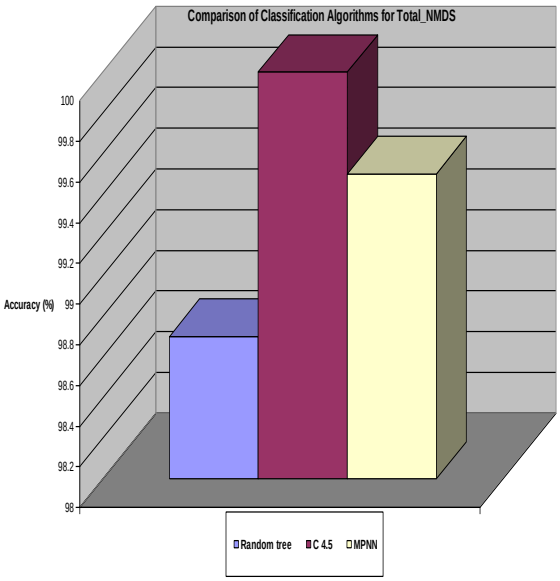

Fig.5 C 4.5 classification algorithm produces 100 percent classification accuracy

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 2, Issue 10, October 2013

V. CONCLUSION

In the proposed work of this research, the oral datasets are get form the various diagnostic centers which contains both cancer and non-cancer patients information and collected data is pre-processed for duplicate and missing information and then We have applied many classification algorithms on NMDS dataset and the performance of those algorithms has been analysed. A classification rate of 100% was obtained for C4.5 algorithm and classification rate of 98.7% was obtained for Random tree Algorithm and classification rate of 99.5% was obtained for MPNN and also to separate the cancer and non-cancer of the patient's records. And then finally we conclude that the datasets get very promising results in classifying the oral cancer. We believe that the proposed system can be very obliging to the doctors for their second outlook for their final decision. In the future Innovation in diagnostic features of tumors may play a central role in development of efficient treatment methods for Oral cancer affected patients. Also shall involve applying image processing technique to diagnose and identifying the stages, and treatments using image datasets.

REFERENCES

[1] S. Warnakulasuriya, "Global epidemiology of oral and oropharyngeal cancer", April-May 2009.
[2] I. Van Der Waal and R. De Bree, "Second primary tumours in oral cancer", June 2010.
[3] A. Jemal, F. Bray, M.M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics", March-April 2011.
[4] www.cancer.org/clinicaltrials. Viewed 28-August- 2013
[5] (www.nccn.org) Viewed 25-August- 2013
[6] Nikhil Sureshkumar Gadewal, Surekha Mahesh Zingde, "Database and interaction network of genes involved in oral cancer", (2011).
[7] Tasnuba Jesmin et.al. "Brain Cancer Risk Prediction Tool Using Data Mining", January 2013
[8] Shantakumar B. Patil and Dr. Y.S. Kumaraswamy., "Extraction of Significant Patterns from Heart Disease Warehouse for Heart Attack Prediction", February 2009.
[9] A.Sudha "Utilization of Data mining Approaches for Prediction of Life Threatening Diseases Survivability", March 2012.
[10] K.Balachandran, Dr. R.Anitha "Supervised Learning Processing Techniques for Pre-Diagnosis of Lung Cancer Disease", 2010.
[11]Hemant palivela*surve .et.al. "On Mining Techniques for Breast Cancer Related Data", 2012*
[12] Arihito Endo et. al., "Comparison of Seven Algorithms to Predict Breast Cancer Survival. Biomedical", 2008.