Using Supervised Learning to Predict Unconscious Oral Cancer

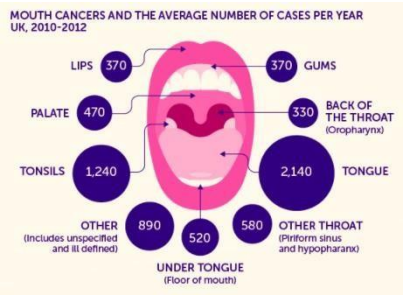
Sonali Pradhan
College of Engineering Bhubaneswar, Biju Pattnaik University of Technology, Odisha, India

Abstract**—** *Since the beginning of time, the disease of cancer has been incurable and intimidating. However owing to the tremendous advancement in the field of technology, it is smoothly curable provided detected in the earliest of stage possible. Precisely considering oral cancer, it is the phenomenon of exponential increase in the number of cells which in turn starts damaging the surrounding and neighboring cells. In spite of the availability of supremely advanced radiation therapy and chemotherapy, the death rate prevailing is very disappointing and increasing. However an early prediction of the same might help to curb this problem. In order to provide with a substantial solution for the same, we propose to perform a comparative analysis of the supervised learning techniques under the domain of machine learning using the accuracy and time complexity approach to design an effective model using the considered data set of the victim to help predict the unconscious cancer in a user so that he/she can work towards the appropriate line of treatment and also make the suggested lifestyle changes. The aim of this paper is to act as a detailed guide for all to develop a system on similar guideline.*

Index Terms**—Supervised Learning, Machine Learning, Oral Cancer, Prediction Model, accuracy, time complexity**.

I. Introduction

Cancer is one of the deadliest diseases such that it ends up claiming millions of lives every single year all around the globe and amongst it, oral cancer is one such sub-type that is mostly triggered due to a few careless day- to-day activities of us human beings and for which we could have a control over. Unfortunately, every year approximately 2,00,000 deaths worldwide and 46,000 deaths particularly in India account for oral cancer. Statistically unlike other types, oral cancer is visible in



the earlier stages on the surface of the mouth in the form of blisters, soft white spots, surfaces getting swollen and extremely red in colour, difficulty in swallowing, excruciating pain in the throat and the mouth region and so on. Good dental or oral care is important to maintaining healthy teeth, gums and tongue. Oral problems, including bad breath, dry mouth, canker or cold sores, TMD, tooth decay, or thrush are all treatable with proper diagnosis and care. Oral cancer can affect any area of the oral cavity including the lips, gum tissues, tongue, cheeklining and the hard and soft palate.

Fig. 1. Affected areas in oral cancer

The following are the precise, elaborate and accurate symp- toms of the oral cancer, which the users need to pay attention to:

1. A sore or blister in your mouth or on your lip thatdoes not heal after two weeks.
2. Lesion on the tongue or tonsil.
3. White and red patches in the mouth or lips that does not heal.
4. Bleeding from the mouth that is unrelated to an injury.
5. Change in the way teeth fit together, including how dentures fit or loose teeth because of jaw swelling or pain.
6. Difficulty swallowing, chewing, speaking, or moving the tongue.

If any of these uneasy symptoms are experienced by the user which could be confused by the user for some regular uneasiness, the computer trained models could help the userto actually verify the probability of actually succumbing to cancer and it would help the use to begin the effective directionof treatment the earliest.

According to the research performed by the multiple health organizations, the factors affecting and supporting the occurrence of the oral cancer are as follows:

1. Consumption of alcohol
2. Consumption of tobacco
3. Gender
4. Age
5. Poor nutrition
6. Immunity deficiencies
7. Viral infections

Owing to all of the factors stated above the user maybe have a larger probability to succumb to the oral cancer. Therefore it also stands important to apprise the user regarding the smaller and a few basic changes he/she could implement in order to have a better overall oral health.

Machine Learning: It is a methodology that includes designing of a model that continues to learn and teach and improvise itself, without any human intervention. There are 3 types of Machine learning techniques:

1. Supervised Learning: Here, the dataset is structured and there exist a certain fixed set of outputs for a fixed set of inputs.
2. Unsupervised Learning: Here, the data available is not structured, however, the output is discovered through a pattern.
3. Reinforcement Learning: A computer program interacts with a dynamic environment in which it must perform a

certain goal.

II. Literature Review

Arushi Tetarbe and Tanushri Choudhury use WEKA Knowl- edge Explorer, a user-friendly GUI, which harness the feature of WEKA software. They have used one more interface in WEKA which has two methods - Explorer interface that devices the statistical knowledge and inference Experimenter interface that analyze the data efficiently by using training and test sets. J48, Random Tree, Naive Bayes, and REP Tree are some of the algorithms that are used. [1]. The survey of machine learning-based approaches was explored to understand the basic application of machine learning inbiomedical research and came across different algorithms and few observations regarding cancer types and attributes [2]. Madhura V, Meghana Nagaraju with their companion survey different reports to study oral cancer detection using machine learning [3]. They then use classification rules for prediction and association rules for showing the co-dependence amongst the attributes. It then uses the apriori algorithm in order to select the frequent itemsets and form the association rule using a bottom-up approach i.e a breadth-first search and a hash to count the items efficiently [4]. The deep learning survival prediction report shows different approaches using the Cox Proportional Hazard regression method and the Random Survival Forest method [5]. Sandhya. N. Dhage primarily provides information regarding various techniques present in the domain of machine learning dividing the approaches into supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning[6Mrs. R. Vidhu, Mrs. S. Kiruthika uses a combination of a genetic algorithm and apriori algorithm as a new feature selection method for better results [11]. It uses association rule mining which is applied to search the hidden relationship among the attributes. A soft computing technique was required for better prediction and understandability of oral cancer in an earlier stage. Zahraa Naser Shah Weli conducted a review of the last ten years to understand the machine learning techniques and its success in the prediction of different types of cancer [12]. Konstantina Kourou, George Rigas gives the idea of dynamic Bayesian networks for the prediction of oral cancer [13]. N.Anitha, K.Jamberi comes up with a more accurate algorithm for the diagnosis and prognosis of oral cancer using a classification algorithm [14]. SHARMA and OM give a detailed version for prediction in terms of age, gender, and socioeconomic status [15]. However, when we conducted a detailed study of the researched material available and is when we encountered the fact that these materials lagged in performing and researching on the available methods by any method and also, it did not explain how can one build an entire model or a system to perform such predictive analysis.

III. Proposed Methodology

The system we intend to build, follows a very simple procedure of taking a few simple inputs from the user based on his/her day to day activities, life style, personal physical details and so on. All of this data collected than shall be tested on the basis of the trained model and the appropriate results shall be generated. The result shall depict the degreeof severity of the cancer the user is prone to. The lifestyle and personal physical attributes causing the severity so that the user can make the changes accordingly and the correct line of treatment could be pursued. From the attribute values gathered a suitable machine learning algorithm shall be processed to find the relevant patterns.

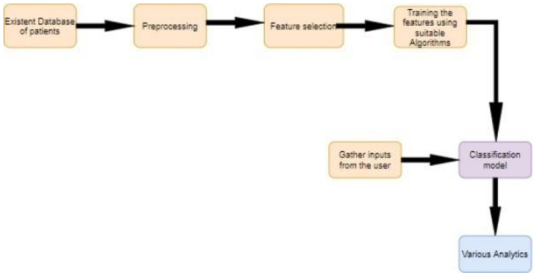


Fig. 3. Usecase diagram of the application

place in training and testing parts, wherein a suitable machine learning algorithm is applied to discover useful patterns and retrieving the same, eventually leading to the construction of a classification model, to which we can supply the acquired inputs.[7]

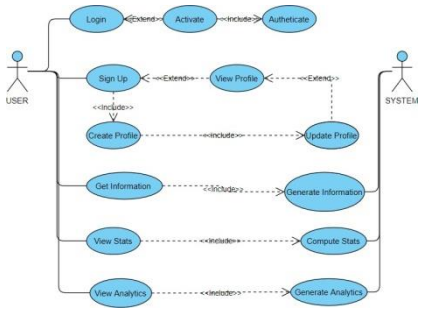


Fig. 2. Flowchart indicating the process model

Above given is the usecase diagram for the system approach we have designed. The usecase diagram tells us how shall the system work. The steps for the same are as follows:

Step 1: If the user has already registered with the system can login in with the valid credentials and then carry on with the relevant objectives.

Step 2: If the person is not registered already, he/she can create an account, sign in and then perform various objectives relatedto the same.

Step 3: The user can view his/her profile with respect to different analytics related to their oral health that might the cause of the disease and also the actual chances of them gettingoral cancer.

The steps that need to be followed and executed The flow of the procedure to be followed in order to build such prediction model is given above. According to chart, initially after taking the input, classification is performed using thebest suitable algorithm subsequently leading to performanceof feature selection.

RESULTS AND DISCUSSIONS

Give below are the steps that we have performed and executed as per the flow chart.

A. Existing Database

The data set obtained for the particular projects consists of 20 attributes of two types, namely the day to day lifestyle related attributes like the food intake and other personal habits and other ones being the bodily related attributes. A combination of these attributes and the pattern in which they vary, shall help us to put up the analytics we intend to.

Here, the attribute 'level' is nominal in nature, the determiningcolumn which shall contribute in deriving the probability of the patient acquiring the cancer. The rest of the attributesare numeric in nature, with their values ranging in particular domain, as per their nature.

The detailed description of the dataset is as follows:

1. Patient ID: This attribute shall uniquely be able to identify the user and the same shall help the application to explicitly revive the records of the particular user. This happens to be a auto-generated feature.

2. Age: The age of the user has a little or no role in determining the probability of cancer, however the age can have an indirect affect on the other values of the user. Therefore, age is takenas one of the attributes as input from the user.

3. Gender: The input is taken from the user for the gender of theirs. We have included two choices for the user to choose from. The user can either be male or female.

4.Tobacco: This is a life style based input. This input is taken from the users on the basis of their tobacco consumption frequency. The scale of this attribute ranges from 1 to 7, where in these numbers represent the number days in a week the users consume tobacco. Where in '1' indicates the consumption being once in a week and '7' indicates the consumption being 7 days in a week.Alcohol-consumption: This is again a life style based input.This input is taken from the users on the basis of their alcohol consumption frequency. The scale of this attribute ranges from1 to 7, where in these numbers represent the number days in a week the users consume alcohol. Where in '1' indicates the consumption being once in a week and '7' indicates the consumption being 7 days in a week.

5. Viral-Infection: This attribute is health based. This indicates how vulnerable the user is to viral infections. The users have4 options to choose from. Namely, 'none', 'rare', 'frequent', 'extreme'. these levels are decided on the frequency of theuser contracting viral infections.

6. Swollen-Tonsil: This input could be given by the users by choosing from two given choices, i.e. either 'yes' or 'no'. Implying if or not the user is experiencing any discomfort atall regarding the swollen tonsil.

7. Genetic-Risk: The genetic risk could be evaluated by the user by choosing from three choices that indicate the historyof cancer in the users' previous generations.

8. Bleeding-Mouth: This attribute is a symptom attribute that indicates the severity of the bleeding experienced by the user. The three types of inputs expected are, 'spotting', 'moderate' and 'extreme'.

9. Balanced-Diet: This is a life style based input. This

inputis taken from the users on the basis of their frequency of consumption of healthy balanced food. The scale of this attribute ranges from 1 to 7, where in these numbers represent the number days in a week the users consume a healthy diet. Where in '1' indicates the consumption being once in a week and '7' indicates the consumption being 7 days in a week.

10. Obesity: This attribute is calculated on the basis of the users' 'BMI'. Depending upon the users' height and weight, the users are divided into 7 categories. With 1 being the lowestBMI and 7 being the highest.
11. Smoking: This is a life style based input. This input is taken from the users on the basis of their smoking frequency. The scale of this attribute ranges from 1 to 7, where in these numbers represent the number days in a week the users smoke up. Where in '1' indicates the consumption being once in a week and '7' indicates the consumption being 7 days in a week.
12. Passive-Smoker: This input could be given by the usersby choosing from two given choices, i.e. either 'yes' or 'no'. Implying if or not the user is indulging into any kind of passivesmoking activity.
13. Red-Spots: This attribute is health based. This indicates how dense spotting is the user facing. The users have 4 options to choose from. Namely, 'none', 'rare', 'frequent', 'extreme'. these levels are decided on the density of the red spots concentration.
14. Coughing-Blood: This is again a health based attribute that indicates the frequency of the user experiencing blood in their cough. The users have 4 options to choose from. Namely,'none', 'rare', 'frequent', 'extreme'.
15. Fatigue: This attribute indicates the extent of fatigue the user has experienced. It could be 'none', the 'regular' and the fatal 'unexplained' one.
16. Weight Loss: This input could be given by the users by choosing from two given choices, i.e. either 'yes' or 'no'. Implying if or not the user is experiencing any unusual weight loss at all.
17. Swallowing-Difficulty: This attribute is health based. This indicates how difficult is the user finding it to swallow stuff. The users have 4 options to choose from. Namely, 'none', 'rare', 'frequent', 'extreme'. these levels are decided on the difficulty of swallowing.
18. Dry-Cough: This input could be given by the users by choosing from two given choices, i.e. either 'yes' or 'no'. Implying if or not the user is experiencing any dry cough.
19. Level: This the final column on which the prediction modelshall be built. This is not taken as an input from the user. This has 3 values, namely, 'Low', 'High', 'Medium'. These levels shall be decided on the basis of the pattern observed in the rest of the attributes mentioned above.

B. Data - Pre-processing

This step involves the cleaning the data available for discrepancy, scaling it and making it ready for actually building the model. Sometimes, it might happen so that from the poolof the data available, some of the values might be missingor inherently incorrect or so. This type of data disrupts the functioning of the model and hinders the accuracy of the model. Therefore, elimination of the same is essential. And exactly the same is achieved in this step. For the dataset spokenabove, we have carried out the following pre-processing steps that were needed:

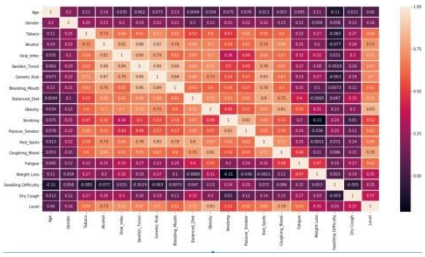
1. Null and missing data: For a few attributes, a few tuples had either null or missing values, in order to deal with this,we filled the missing spaces with the mean of the rest of the values of the other tuples.
2. Encoding the columns: In order to deal with the categorical data, we used the label encoder system in order to convert the data into respective data into the corresponding labels in order to facilitate easier training of the model.
3. Scaling the attributes: For the dataset we had a few columns that had a huge difference in their range of values which could cause an issue in training the model, therefore, we scaled a few attributes in correspondence to each other.

C. Feature Selection

Feature selection is the process, in which we select a subset of features from all the existing features depending upon their co-relation to the final and deciding column. In order to carry this step out, we implemented the same using two methods:

1. Plotted a 'Correlation Matrix'.
2. Calculated the 'Covariance' values.

1. Correlation Graph: It a technique of plotting attributes against each other, where in, owing to the correlation value of the attributes, we can be able to decipher the correlation between different attributes available in the dataset and that helps us understand how these attributes are related to each other and to what extent.



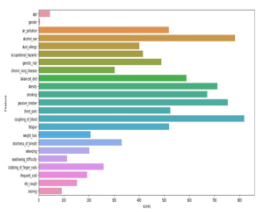


Fig. 4. A matrix depicting the correlation between difference attributes

TABLE I
ATTRIBUTES AND THEIR CORRELATION VALUES

Attribute	Correlation Value
Tobacco	0.64
Alcohol	0.73
Viral-Infection	0.71
Swollen-Tonsil	0.67
Genetic-Risk	0.70
Bleeding-Mouth	0.61
Balanced-Diet	0.71
Obesity	0.83
Smoking	0.52
Passive-Smoking	0.62
Red-Spots	0.65
Coughing-Blood	0.78

The graph plotted above is an attempt to derive the relation between the 20 attributes we have covered in our dataset. The above drawn figure is a 'Heat Map', that allows to understand the 'heated region', the area in orange, where the correlationis thickening.

2. Covariance Values: Covariance values depict how the values from the target column vary with the values of thethe attributes that go into the making of the model. It depicts the linear variance of the output variables with the intended features. From the entire set of data we have, the attributes with the highest values of covariance in their descending order are given in the table. Here we have used the 'Chi Square' formula to obtain these values of covariance. After implementing both these methods an intersection of attributes from both these outputs are selected.

D. Training the model

In order to train the model for producing the relevant predictions, we can opt to two ways; namely a 'Classification model' or a 'Regression model'. However, for our work we have chosen to train our model we have chosen to build a classification model, since, our deciding column has distinct categorial values. Therefore, we have opted for a natural choice of a classification model to classify the final column('Level') into either 'High', 'Medium' or 'Low'. For building the classification model, we focused on not only achieving a solution but also on obtaining an

Fig. 5. A barplot depicting the attributes and their Chi Sq. Covarinace scores

TABLE II
ATTRIBUTES AND THEIR COVARIANCE VALUES

Attribute	Covariance Value
Tobacco	818.668884
alcohol-use	781.909841
passive-smoker	752.959791
obesity	712.087562
smoking	671.006253
balanced-diet	588.933743
chest-pain	524.489521
fatigue	518.900446
air-pollution	518.631533
genetic-risk	488.649726
occupational-hazards	415.685654
dust-allergy	401.040883
shortness-of-breath	330.880709
chronic-lung-disease	302.396157
clubbing-of-finger-nails	257.907679
weight-loss	206.666563
wheezing	201.426189
frequent-cold	192.713276
dry-cough	152.029547
swallowing-difficulty	113.074249

optimized solution. In order to achieve the optimized resultfor our model, it was essential for us to use the best fit of the algorithm, for which we carried out the comparative studies of the algorithms available for supervised learning. For starters, the supervised learning technique in the machine learning methodology deals with developing a model where for a particular set of inputs, there exists a fixed set of outputs corresponding to the inputs. In order to achieve this, we have studied four supervised algorithms namely, K Nearest Neighbours, Decision Trees, Naive Bayes and Random Forest.[8]

In order to train the model for the supervised learning, wehave divided the dataset into the 80:20 ratio. Where in, we have used 80 percent of the data[800 tuples], to train the model, and the remaining 20 percent data[200 tuples] to test the dataset for. In order to select the one that shall suit the best given the selected dataset and provide the most accurate result, we have implemented the accuracy scores understood through the 'Confusion Matrix'. A 'Confusion Matrix' is a matrix plot that tells us the number of instances correctly and incorrectly classified by a particular algorithm.

The confusion matrix for Naive Bayes algorithm is as follows:

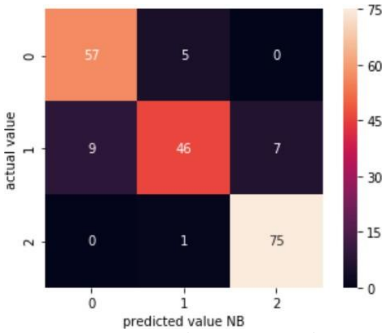


Fig. 6. Confusion Matrix for Naive Bayes Algorithm

For the Naive Bayes algorithm:True Positives: 178/200
False Positives: 22/200 Percentage Accuracy: 89 percent
For the Decision Tree algorithm:True Positives: 200/200
False Positives: 0/200
Percentage Accuracy: 100 percent

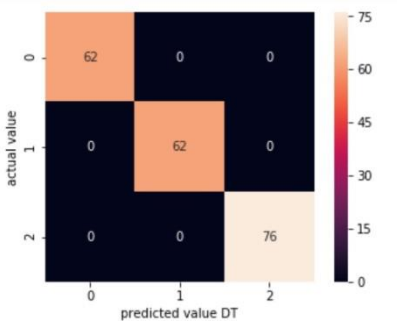


Fig. 8. Confusion Matrix for Decision Tree Algorithm

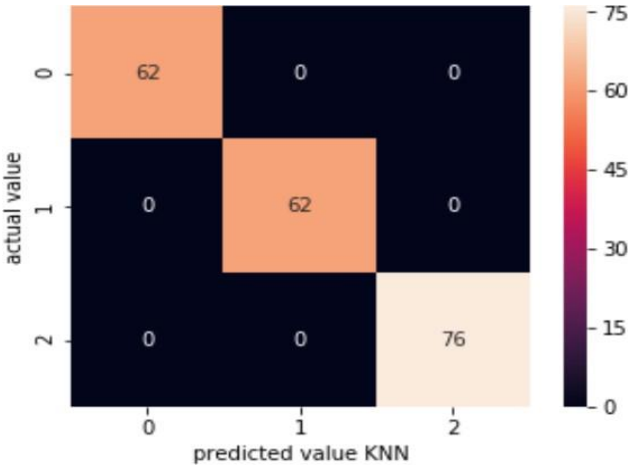


Fig. 7. Confusion Matrix for KNN Algorithm

The confusion matrix for KNN algorithm is as above:For the KNN algorithm:
True Positives: 200/200False Positives: 0/200
Percentage Accuracy: 100 percent

The confusion matrix for Decision Tree algorithm is asabove:

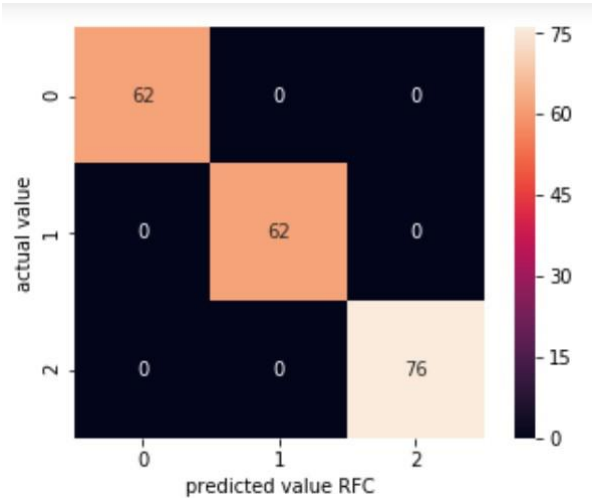


Fig. 9. Confusion Matrix for Random Forest Classifier Algorithm

The confusion matrix for Random Forest Classifieralgorithm is as above:
For the Random Forest Classifier algorithm:True Positives: 200/200
False Positives: 0/200
Percentage Accuracy: 100 percent

From our analysis we understood that except for the Naive

Algorithms	Training	Prediction
Decision Tree	$O(n^2p)$	O(p)
Random Forest	$O(n^2pn)$	O(pn)
KNN	-	O(np)

TABLE III
ALGORITHMS AND THEIR TRAINING AND PREDICTION COMPLEXITIES

Bayes algorithm, all the other algorithms gave us a 100 percent accuracy. Therefore, in order to select and algorithm from the other three algorithms, we took into consideration the time complexities of the algorithms, in order to get an idea of the amount of time required by the algorithms to render the output. From these studies, we discovered that, the complexities of the algorithms were elicited in the table:

From the table it is evidently clear that the Decision Tree algorithms needs the least of the time period to render and produce the output.

Therefore, based on the accuracy(100 percent) and time complexity

$o[n^2p,$

$p]$*wedecidedtouseDecisionTreetotrainourmodel.*

And using these steps, we successfully trained our model.

IV. LIMITATIONS

The limitations of our approach are as follows:

1. We have the scope to research, experiment and comparethe deep learning approach for the model.
2. Obtaining an authentic dataset for this particular domaincould be very challenging.
3. The accuracy of supervised learning models could be varied because of the outliers.
4.

V. CONCLUSION

The above given detailed description of the procedure to be followed is based on the supervised learning methodology, and we selected this method because it was best suited for our dataset; as, the data was labeled. However, as we worked on this project, the advantages and disadvantages of using this approach are summarized as follows:

Advantages: 1. The model provides fast and efficient outputs for the data that is labeled and has a consistent set of outputs for corresponding inputs.

2. In supervised learning, we have an exact idea about the classes of the objects.

Disadvantages: 1. It is not suitable to handle complex tasks.

2. The training requires a lot of computational times.

3. Might not provide accurate answers for the inputs that deviate a lot from the training dataset.

REFERENCES

[1] Arushi Tetarbe, Tanupriya Choudhary, Teoh Yiek Toe, Seema Rawat “Oral Cancer Detection using data mining tool” 2017 IEEE.

[2] Ajay Kumar, Rama Sushil, Arvind Kumar Tiwari “Machine Learning based Approaches for Cancer Prediction: A Survey” 2019.

[3] Madhura V, Meghana Nagaraju, Namana J, Varshini S, Rakshitha R “Survey Paper on Oral Cancer Detection using Machine Learning” 2019.

[4] Madhura V, Meghana Nagaraju, Namana J, Varshini S, Rakshitha R “Oral Cancer Detection Using Machine Learning “ 2019DongWook Kim , Sanghoon Lee, Sunmo Kwon, Woong Nam, In-HoCha Hyung Jun Kim”Deep learning-based survival prediction of oral cancer patients” 2019.

[5] Sandhya N. Dhage “A Review on Early Detection of Oral Cancer using ML Techniques” 2019.

[6] K. Lalithamani, A. Punitha ”Detection of Oral Cancer using Deep Neural Based Adaptive Fuzzy System in data mining techniques” 2019.

[7] Lavanya, Dr. Chandra J “Oral Cancer Analysis Using Machine Learning Techniques” 2019.

[8] Fatihah Mohd, Noor Maizura Mohamad Noor, Zainab Abu Bakar, Zainul Ahmad Rajion “Analysis of Oral Cancer Prediction using Features Selection with Machine Learning” 2015.

[9] Shikha Agrawal, Jitendra Agrawal “Neural Network Techniques for Cancer Prediction: A Survey” 2016.

[10] Mrs. R. Vidhu, Mrs. S. Kiruthika “A New Feature Selection Method for Oral Cancer Using Data Mining Techniques” 2016.

[11] Zahraa Naser Shah Weli “Data Mining in Cancer Diagnosis and Predic- tion: Review about Latest Ten Years” 2020.

[12] Konstantina Kourou, George Rigas, Konstantinos P. Exarchos, Costas Papaloukas and Dimitrios I. Fotiadis “Prediction of Oral Cancer Recur- rence using Dynamic Bayesian Networks”.

[13] N.Anitha, K.Jamberi “Diagnosis and Prognosis of Oral Cancer using classification algorithm with Data Mining Techniques”

[14] Sharma N., Om H. “Using Data Mining For Oral Cancel Risk Stratifi- cation in terms of Age, Gender And Socioeconomic Status” 2013.

[15] Konstantina Kourou a, Themis P. Exarchos a,b, Konstantinos P. Exarchos a, Michalis V. Karamouzis c, Dimitrios I. Fotiadis “Machine learning applications in cancer prognosis and prediction” 2015.