ISSN No. 0976-5697

Volume 3, No. 1, Jan-Feb 2012

International Journal of Advanced Research in Computer Science

CASE STUDY AND REPORT

Available Online at www.ijarcs.info

An Evaluation study of Oral Cancer Detection using Data Mining Classification Techniques

S.Prasanna*
SITE,VIT University,
Vellote,India
sprasanna@vit.ac.in

K.Govinda
SCSE,VIT University,
Vellote, India
kgovinda@vit.ac.in

U.Senthil Kumaran
SITE,VIT University,
Vellote, India
usenthilkumaran@vit.ac.in

Abstract: Cancer detection is one of the important research topics in medical science. Oral cancer is the sixth most common cancer in the world. It is one of the most prevalent cancers in the developing countries of South Asia accounting for one third of the world burden. In India oral cancer is the most common cancer that occurs. Sixty percent of the cancers are advanced by the time they are detected. In this paper we have implemented two techniques such as Naïve Bayesian and Support Vector Machine (SVM) and compared the results to show which technique is the best. Current predictive model design in medical oncology literature is dominated by linear and logistic regression techniques. In IPPSCD (Intelligent Prognosis Prediction System for Cancer Disease) a database of cancer patient performance is constructed. Data mining techniques will be used to analyze the database to predict disease based on causative factors. Both classification and regression algorithms are to be considered

Keywords: SVM, LIF, CCD , HPLC, MPM, SHG.

I. INTRODUCTION

In India, 320,000 new cancer cases are reported (males) every year and oral cancer accounts for 19% of them, while in women it accounts for about 7% of the total of 350,000 new cancer cases [1]. One third of world's oral cancer population resides in Indian subcontinent. This is due to the consumption of tobacco and alcohol. The worldwide annual incidence and mortality due to oral cancer is around 274,000 and 127,000 respectively [2]. Currently, the most widely used screening test for oral cancer is visual inspection of the oral cavity. Hence, need of the hour is a technology that's easy and simple (can be managed by less skilled providers at rural areas) aimed at screening of precancerous lesions so that it has a much wider population reach and acceptance. This proposed system envisages predicting oral cancer at an early stage. The system uses data attributes like person's id, person's name, age, gender, alcohol, smoking, ulcer status to predict oral cancer. Here we used Naïve Bayesian and Support vector machine (SVM) for classification. There are many classification algorithms. But Naïve Bayesian is a very simple, basic classification system and the SVM gives much more accurate results than most of the other classification algorithms [3].

II. OPTICAL SCREENING OF ORAL CANCER

Two different technologies LIF and HPLC-LIF have been developed for the early screening of oral cancer.LIF is being developed for in-vivo screening whereas HPLC-LIF is for in-vitro diagnostics.[4]

A. Laser Induced Fluorescence (LIF) Technology:

A laser beam is transmitted to the tissue through a laser carrying fiber. The tissue emits fluorescence with maximum emission for NADH at 440nm and for collagen at 400nm.Collection fibers in the probe carry the auto fluorescence signal of the tissue to the spectrograph. The fluorescence wavelength components dispersed by the spectrograph are recorded using a Charge Coupled Device (CCD) and the spectrum is recorded in the computer[5].

B. High Performance Liquid Chromatography-LIF:

HPLC-Biotechnology is used to generate protein profiles in the form of chromatogram from human serum and saliva, which can be used to screen oral cancer using the signatures of the unidentified tumor markers embedded in the chromatograms [6].

C. Imaging of human oral cancer using multiphoton microscopy:

TPM is a nonlinear high resolution optical method have been used in a variety of biological imaging applications, Two-photon interactions in MPM result in SHG and two photon excited fluorescence (TPF). The goal of this study is to investigate the application of a multimodal nonlinear Imaging approach. It is based on the integration of Multiphoton microscopy (MPM) with second harmonic generation microscopy (SHGM) for in vivo evaluation of oral tissue microstructure. This study chooses to use two-photon fluorescence and second-harmonic generation microscopy as imaging techniques. A difference in SHG and auto fluorescence images exists between normal and cancerous tongue tissues. Such differences may be used for the in vivo diagnosis of normal and cancerous tongue tissues. This was planed to be extended into the multiphoton characterization of other oral diseases. This was thought to help in the diagnosis and treatment of oral diseases. The paper is organized in the following sequence. A brief literature survey of the oral cancer, etc. was presented in the previous paragraphs

III. NAÏVE BAYESIAN CLASSIFICATION

The naive Bayesian classifier, or simple Bayesian classifier, works as follows[7].

A. Each data sample is represented by an n-dimensional feature vector, X = (x1; x2; : : :; xn), depicting n measurements made on the sample from n attributes, respectively A1;A2; ::;An.

B. Suppose that there are m classes, C1; C2; : : :; Cm. Given an unknown data sample, X (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naive Bayesian classifier assigns an unknown sample X to the class Ci if and only if :

P(Ci/X) > P(Cj/X) for 1 <= j <= m; j != i -- eq(2)

Thus we maximize P(Ci/X). The class Ci for which P(Ci/X) is maximized is called the maximum posteriori hypothesis. By Bayes theorem (Equation (1))[12],

P(Ci/X) =P(X/Ci)P(Ci) / P(X) -- eq(3)

C. As P(X) is constant for all classes, only P(XjCi)P(Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, i.e. P(C1) = P(C2) = : : : = P(Cm), and we would therefore maximize P(XjCi). Otherwise, we maximize P(XjCi)P(Ci). Note that the class prior probabilities may be estimated by P(Ci) = si /s , where si is the number of training samples of class Ci, and s is the total number of training samples.

Given data sets with many attributes, it would be extremely computationally expensive to compute P(XjCi).

D. Given data sets with many attributes, it would be extremely computationally expensive to compute P(XjCi). In order to reduce computation in evaluating P(XjCi), the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample, i.e., that there are no dependence relationships among the attributes. Thus,

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i).$$

eq(4)

The probabilities P(x1/Ci); P(x2/Ci); : : :; P(xn/Ci) can be estimated from the training samples, where:

a. If Ak is categorical, then P(xk/Ci) = sik/si,

where sik is the number of training samples of class Ci having the value xk for Ak, and si is the number of training samples belonging to Ci.

b. If Ak is continuous-valued, then the attribute is assumed to have a Gaussian distribution. Therefore,

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sqrt{2\pi}\sigma_{C_i}} e^{-\frac{(x-\mu_{C_i})^2}{2\sigma^2_{C_i}}},$$

-- eq(4)

where g(xk,μ_{Ci},σ_{Ci}) is the Gaussian (normal)density function for attribute Ak, while μ_{Ci} and σ_{Ci} are the mean and variance respectively given the values for attribute Ak for training samples of class Ci.

E. In order to classify an unknown sample X, P(X/Ci)P(Ci) is evaluated for each class Ci. Sample X is then assigned to the class Ci if and only if :

P(X/Ci)P(Ci) > P(X/Cj)P(Cj) for 1 <=j<=m; j != i. –eq(5)

In other words, it is assigned to the class, Ci, for which P(X/Ci)P(Ci) is the maximum. Bayesian classifiers are also useful in that they provide a theoretical justification for other classifiers which do not explicitly use Bayes theorem.

IV. SUPPORT VECTOR MACHINE

SVM is a supervised learning methods used for classification and regression. In simple words, given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other. Intuitively, an SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. A support vector machine constructs a hyperplane or set of hyper planes in a high or infinite dimensional space, which can be used for classification, regression or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

Our task is to predict whether a test sample belongs to one of two classes. We receive training examples of the form: $\{x_i; y_i\}$, $i = 1,.., n$ and $x_i \epsilon R^d; y_i \epsilon\{-1,+1\}$. We call *{*x*i}* the co-variates or input vectors and *{yi}* the response variables or labels.

We consider a very simple example where the data are in fact linearly separable: i.e. I can draw a straight line $f(x) = wT\,x - b$ such that all cases with $yi = -1$ fall on one side and have $f(xi) < 0$ and cases with $yi = +1$ fall on the other and have $f(xi) > 0$. Given that we have achieved that, we could classify new test cases according to the rule $y_{test} = \text{sign}(x_{test})$.

However, typically there are infinitely many such hyper-planes obtained by small perturbations of a given solution. How do we choose between all these hyper-planes which the solve the separation problem for our training data, but may have different performance on the newly arriving test cases. For instance, we could choose to put the line very close to members of one particular class, say $y = +1$. Intuitively, when test cases arrive we will not make many mistakes on cases that should be classified with $y = +1$, but we will make very easily mistakes on the cases with $y = -1$ (for instance, imagine that a new batch of test cases arrives which are small perturbations of the training data). A sensible thing thus seems to choose the separation line as far away from both y = -1 and y = +1 training cases as we can, i.e. right in the middle.

Geometrically, the vector w is directed orthogonal to the line defined by $w^T x = b$. This can be understood as follows. First take b = 0. Now it is clear that all vectors, x, with vanishing inner product with w satisfy this equation, i.e. all vectors orthogonal to w satisfy this equation. Now translate the hyperplane away from the origin over a vector a. The equation for the plane now becomes: $(x - a)^T w = 0$, i.e. we find that for the offset $b = a^T w$, which is the projection of a onto to the vector w. Without loss of generality we may thus choose a perpendicular to the plane, in which case the length ||a|| = |b| / ||w|| represents the shortest, orthogonal distance between the origin and the hyperplane. We now define 2 more hyperplanes parallel to the separating hyperplane. They represent that planes that cut through the closest training examples on either side. We will call them "support hyper-planes" in the following, because the data-vectors they contain support the plane.

We define the distance between the these hyperplanes and the separating hyperplane to be d_+ and d_- respectively. The *margin*, γ, is defined to be $d_+ + d_-$. Our goal is now to find the separating hyperplane so that the margin is largest, while the separating hyperplane is equidistant from both.

We can write the following equations for the support hyperplanes:

$wT\,x = b + \delta$ -- (1)

$w^T x = b - \delta$ --(2)

We now note that we have over-parameterized the problem: if we scale w,b and δ by a constant factor α, the equations for x are still satisfied. To remove this ambiguity we will require that δ = 1, this sets the scale of the problem, i.e. if we measure distance in meters or millimeters.We can now also compute the values for

$d_+ = (||b+1| - |b||) / ||w|| = 1/||w||$

(this is only true if b *not* ϵ (-1,0) since the origin doesn't fall in between the hyperplanes in that case. If b ϵ (-1, 0) you should use $d_+ = (||b + 1| + |b||) / ||w|| = 1/ ||w||$). Hence the margin is equal to twice that value: $\gamma = 2 / ||w||$. With the above definition of the support planes we can write down the following constraint that any solution must satisfy,

$w^T x_i - b <= -1$ ¥ $yi = -1$ (3)

$w^T x_i - b >= +1$ ¥ $yi = +1$ (4)

or in one equation,

$yi(w^T x_i - b) - 1 >= 0$ (5)

We now formulate the primal problem of the SVM:

Minimize $1/2||w||^2$

subject to $y_i(wT\,x_i - b) - 1 >= 0$ ¥i (6)

Thus, we maximize the margin, subject to the constraints that all training cases fall on either side of the support hyper-planes. The data-cases that lie on the hyperplane are called support vectors, since they support the hyper-planes and hence determine the solution to the problem.

The primal problem can be solved by a quadratic program. However, it is not ready to be kernelised, because its dependence is not only on inner products between data-vectors. Hence, we transform to the dual formulation by first writing the problem using a Lagrangian,

$$\mathcal{L}(\mathrm{w}, b, \alpha) = \frac{1}{2}||\mathrm{w}||^2 - \sum_{i=1}^{N} \alpha_i \left[y_i(\mathrm{w}^T \mathrm{x}_i - b) - 1 \right]$$

The solution that minimizes the primal problem subject to the constraints is given by $\min_w \max\alpha\ L(w;\alpha)$, i.e. a saddle point problem. When the original objective-function is convex, (and only then), we can interchange the minimization and maximization. Doing that, we find that we can find the condition on w that must hold at the saddle point we are solving for. This is done by taking derivatives wrt w and b and solving

$$\mathbf{w} - \sum_i \alpha_i y_i \mathbf{x}_i = 0 \quad \Rightarrow \quad \mathbf{w}^* = \sum_i \alpha_i y_i \mathbf{x}_i$$

$$\sum_i \alpha_i y_i = 0$$

Inserting this back into the Lagrangian we obtain what is known as the dual problem,

$$\text{maximize} \quad \mathcal{L}_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{subject to} \quad \sum_i \alpha_i y_i = 0$$

$$\alpha_i \geq 0 \ \forall i$$

The dual formulation of the problem is also a quadratic program, but note that the number of variables, α_i in this problem is equal to the number of data-cases, N. The crucial point is however, that this problem *only depends on* x*i through the inner-product* $x^T_{i.}\, x_j$. This is readily kernelised through the substitution $x^T_{i.}\ xj$ -> $k(xi;\ xj)$. This is a recurrent theme: the dual problem lends itself to kernelisation, while the primal problem is not kernelised.

V. PROPOSED WORK

The proposed system architecture consists of the following Phases (i) Data collection, (ii) preprocessing. (iii) implementation of naïve bayes algorithm. (iv)Implementation of SVM. (v) comparison of naïve bayes and SVM. (vi) cancer prediction.

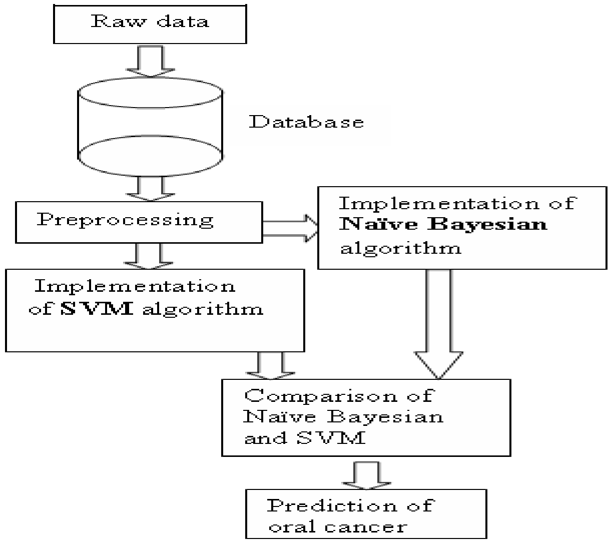
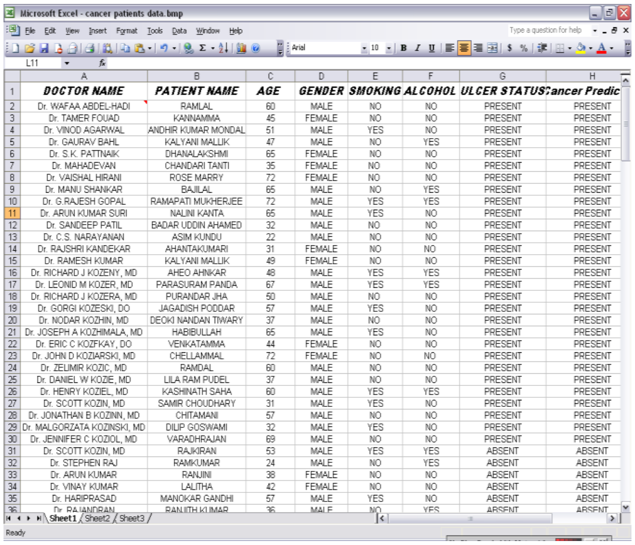

Raw data
Database
Preprocessing
Implementation of Naïve Bayesian algorithm
Implementation of SVM algorithm
Comparison of Naïve Bayesian and SVM
Prediction of oral cancer

Figure1. System Architecture

A. Data Collection and Preprocessin:

In data collection module the following data like doctor name, doctor_id, patient_id , patient name, age, smoking, and alcohol and ulcer status are collected and stored in the database Oracle 10g.once the data is stored in the database it is just like a raw data and it is preprocessed for applying the machine learning algorithms in Clementine tool to predict the presence of cancer.in the preprocessing stage the data is stored in excel sheet as shown in Fig3 without any missing values and then stored in text file with delimiters as shown in Fig4.



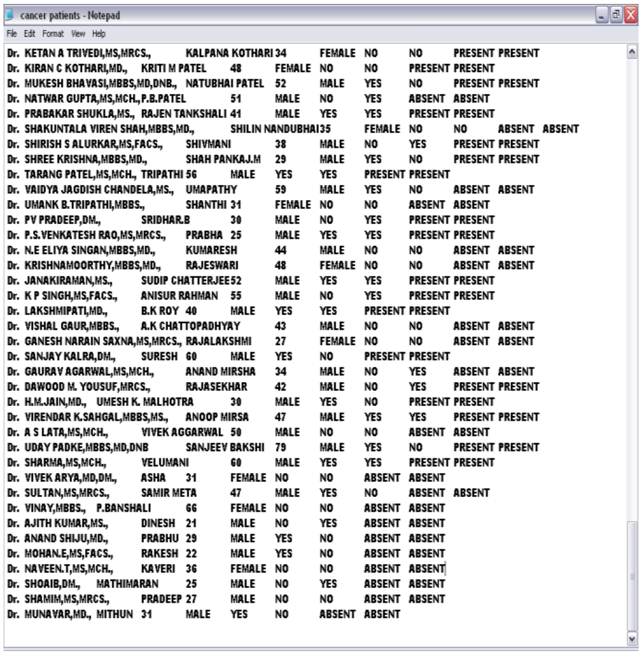
Microsoft Excel - cancer patients data.bmp

	A	B	C	D	E	F	G	H
1	DOCTOR NAME	PATIENT NAME	AGE	GENDER	SMOKING	ALCOHOL	ULCER STATUS	?ancer Predic
2	Dr. WAFAA ABDEL-HADI	RAMLAL	60	MALE	NO	NO	PRESENT	PRESENT
3	Dr. TAMER FOUAD	KANNAMMA	45	FEMALE	NO	NO	PRESENT	PRESENT
4	Dr. VINOD AGARWAL	ANDHIR KUMAR MONDAL	51	MALE	YES	NO	PRESENT	PRESENT
5	Dr. GAURAV BAHL	KALYANI MALLIK	47	MALE	NO	YES	PRESENT	PRESENT
6	Dr. S.K. PATTNAIK	DHANALAKSHMI	65	FEMALE	NO	NO	PRESENT	PRESENT
7	Dr. MAHADEVAN	CHANDARI TANTI	35	FEMALE	NO	NO	PRESENT	PRESENT
8	Dr. VAISHAL HIRANI	ROSE MARRY	72	FEMALE	NO	NO	PRESENT	PRESENT
9	Dr. MANU SHANKAR	BAJILAL	65	MALE	NO	YES	PRESENT	PRESENT
10	Dr. G.RAJESH GOPAL	RAMAPATI MUKHERJEE	72	MALE	YES	YES	PRESENT	PRESENT
11	Dr. ARUN KUMAR SURI	NALINI KANTA	65	MALE	YES	NO	PRESENT	PRESENT
12	Dr. SANDEEP PATIL	BADAR UDDIN AHAMED	32	MALE	NO	NO	PRESENT	PRESENT
13	Dr. C.S. NARAYANAN	ASIM KUNDU	22	MALE	NO	NO	PRESENT	PRESENT
14	Dr. RAJSHRI KANDEKAR	AHANTAKUMARI	31	FEMALE	NO	NO	PRESENT	PRESENT
15	Dr. RAMESH KUMAR	KALYANI MALLIK	49	FEMALE	NO	NO	PRESENT	PRESENT
16	Dr. RICHARD J KOZENY, MD	AHEO AHNKAR	48	MALE	YES	YES	PRESENT	PRESENT
17	Dr. LEONID M KOZER, MD	PARASURAM PANDA	67	MALE	YES	YES	PRESENT	PRESENT
18	Dr. RICHARD J KOZERA, MD	PURANDAR JHA	50	MALE	NO	NO	PRESENT	PRESENT
19	Dr. GORGI KOZESKI, DO	JAGADISH PODDAR	57	MALE	YES	NO	PRESENT	PRESENT
20	Dr. NODAR KOZHIN, MD	DEOKI NANDAN TIWARY	37	MALE	NO	NO	PRESENT	PRESENT
21	Dr. JOSEPH A KOZHIMALA, MD	HABIBULLAH	65	MALE	YES	NO	PRESENT	PRESENT
22	Dr. ERIC C KOZFKAY, DO	VENKATAMMA	44	FEMALE	NO	NO	PRESENT	PRESENT
23	Dr. JOHN D KOZIARSKI, MD	CHELLAMMAL	72	FEMALE	NO	NO	PRESENT	PRESENT
24	Dr. ZELIMIR KOZIC, MD	RAMDAL	60	MALE	NO	NO	PRESENT	PRESENT
25	Dr. DANIEL W KOZIE, MD	LILA RAM PUDEL	37	MALE	NO	NO	PRESENT	PRESENT
26	Dr. HENRY KOZIEL, MD	KASHINATH SAHA	60	MALE	YES	YES	PRESENT	PRESENT
27	Dr. SCOTT KOZIN, MD	SAMIR CHOUDHARY	31	MALE	YES	NO	PRESENT	PRESENT
28	Dr. JONATHAN B KOZINN, MD	CHITAMANI	57	MALE	NO	NO	PRESENT	PRESENT
29	Dr. MALGORZATA KOZINSKI, MD	DILIP GOSWAMI	32	MALE	YES	NO	PRESENT	PRESENT
30	Dr. JENNIFER C KOZIOL, MD	VARADHRAJAN	69	MALE	NO	NO	PRESENT	PRESENT
31	Dr. SCOTT KOZIN, MD	RAJKIRAN	53	MALE	YES	YES	ABSENT	ABSENT
32	Dr. STEPHEN RAJ	RAMKUMAR	24	MALE	NO	YES	ABSENT	ABSENT
33	Dr. ARUN KUMAR	RANJINI	38	FEMALE	NO	NO	ABSENT	ABSENT
34	Dr. VINAY KUMAR	LALITHA	42	FEMALE	NO	NO	ABSENT	ABSENT
35	Dr. HARIPRASAD	MANOKAR GANDHI	57	MALE	YES	NO	ABSENT	ABSENT
36	Dr. RAJANDRAN	RANJITH KUMAR	36	MALE	NO	YES	ABSENT	ABSENT

Sheet1 / Sheet2 / Sheet3

Ready

Figure3. Sample Data



cancer patients - Notepad

```
Dr.  KETAN A TRIVEDI,MS,MRCS.,        KALPANA KOTHARI 34        FEMALE  NO      NO      PRESENT PRESENT
Dr.  KIRAN C KOTHARI,MD.,    KRITI M PATEL      48        FEMALE  NO      NO      PRESENT PRESENT
Dr.  MUKESH BHAVASI,MBBS,MD,DNB.,     NATUBHAI PATEL  52        MALE    YES     NO      PRESENT PRESENT
Dr.  NATWAR GUPTA,MS,MCH.,P.B.PATEL             51        MALE    NO      YES     ABSENT  ABSENT
Dr.  PRABAKAR SHUKLA,MS.,    RAJEN TANKSHALI 41        MALE    YES     YES     PRESENT PRESENT
Dr.  SHAKUNTALA VIREN SHAH,MBBS,MD.,          SHILIN NANDUBHAI35        FEMALE  NO      NO      ABSENT  ABSENT
Dr.  SHIRISH S ALURKAR,MS,FACS.,      SHIVMANI          38        MALE    NO      YES     PRESENT PRESENT
Dr.  SHREE KRISHNA,MBBS,MD.,          SHAH PANKAJ.M     29        MALE    YES     NO      PRESENT PRESENT
Dr.  TARANG PATEL,MS,MCH.,   TRIPATHI 56        MALE    YES     YES     PRESENT PRESENT
Dr.  VAIDYA JAGDISH CHANDELA,MS.,     UMAPATHY          59        MALE    YES     NO      ABSENT  ABSENT
Dr.  UMANK B.TRIPATHI,MBBS.,          SHANTHI 31        FEMALE  NO      NO      ABSENT  ABSENT
Dr.  PV PRADEEP,DM.,         SRIDHAR.B          30        MALE    NO      YES     PRESENT PRESENT
Dr.  P.S.VENKATESH RAO,MS,MRCS.,      PRABHA  25        MALE    YES     YES     PRESENT PRESENT
Dr.  N.E ELIYA SINGAN,MBBS,MD.,       KUMARESH          44        MALE    NO      NO      ABSENT  ABSENT
Dr.  KRISHNAMOORTHY,MBBS,MD.,         RAJESWARI         48        FEMALE  NO      NO      ABSENT  ABSENT
Dr.  JANAKIRAMAN,MS.,        SUDIP CHATTERJEE52        MALE    YES     YES     PRESENT PRESENT
Dr.  K P SINGH,MS,FACS.,     ANISUR RAHMAN      55        MALE    NO      YES     PRESENT PRESENT
Dr.  LAKSHMIPATI,MD.,        B.K ROY  40        MALE    YES     YES     PRESENT PRESENT
Dr.  VISHAL GAUR,MBBS.,      A.K CHATTOPADHYAY          43        MALE    NO      NO      ABSENT  ABSENT
Dr.  GANESH NARAIN SAXNA,MS,MRCS.,    RAJALAKSHMI       27        FEMALE  NO      NO      ABSENT  ABSENT
Dr.  SANJAY KALRA,DM.,       SURESH   60        MALE    YES     NO      PRESENT PRESENT
Dr.  GAURAV AGARWAL,MS,MCH.,          ANAND MIRSHA      34        MALE    NO      YES     ABSENT  ABSENT
Dr.  DAWOOD M. YOUSUF,MRCS.,          RAJASEKHAR        42        MALE    NO      YES     PRESENT PRESENT
Dr.  H.M.JAIN,MD.,  UMESH K. MALHOTRA           30        MALE    YES     NO      PRESENT PRESENT
Dr.  VIRENDAR K.SAHGAL,MBBS,MS.,      ANOOP MIRSA       47        MALE    YES     YES     PRESENT PRESENT
Dr.  A S LATA,MS,MCH.,       VIVEK AGGARWAL  50        MALE    NO      NO      ABSENT  ABSENT
Dr.  UDAY PADKE,MBBS,MD,DNB           SANJEEV BAKSHI    79        MALE    YES     NO      PRESENT PRESENT
Dr.  SHARMA,MS,MCH.,         VELUMANI           60        MALE    YES     YES     PRESENT PRESENT
Dr.  VIVEK ARYA,MD,DM.,      ASHA     31        FEMALE  NO      NO      ABSENT  ABSENT
Dr.  SULTAN,MS,MRCS.,        SAMIR META         47        MALE    YES     NO      ABSENT  ABSENT
Dr.  VINAY,MBBS.,   P.BANSHALI          66        FEMALE  NO      NO      ABSENT  ABSENT
Dr.  AJITH KUMAR,MS.,        DINESH   21        MALE    NO      YES     ABSENT  ABSENT
Dr.  ANAND SHIJU,MD.,        PRABHU   29        MALE    YES     NO      ABSENT  ABSENT
Dr.  MOHAN.E,MS,FACS.,       RAKESH   22        MALE    YES     NO      ABSENT  ABSENT
Dr.  NAVEEN.T,MS,MCH.,       KAVERI   36        FEMALE  NO      NO      ABSENT  ABSENT
Dr.  SHOAIB,DM.,    MATHIMARAN          25        MALE    NO      YES     ABSENT  ABSENT
Dr.  SHAMIM,MS,MRCS.,        PRADEEP  27        MALE    NO      NO      ABSENT  ABSENT
Dr.  MUNAVAR,MD.,   MITHUN   31        MALE    YES     NO      ABSENT  ABSENT
```

Figure4. Preprocessed Data

B. Implementation of Naïve Bayes Algorithm:

The Naïve Bayesian Network, the target prediction of oral cancer is calculated by the attributes displayed in Fig2 and the result is shown in Fig5.

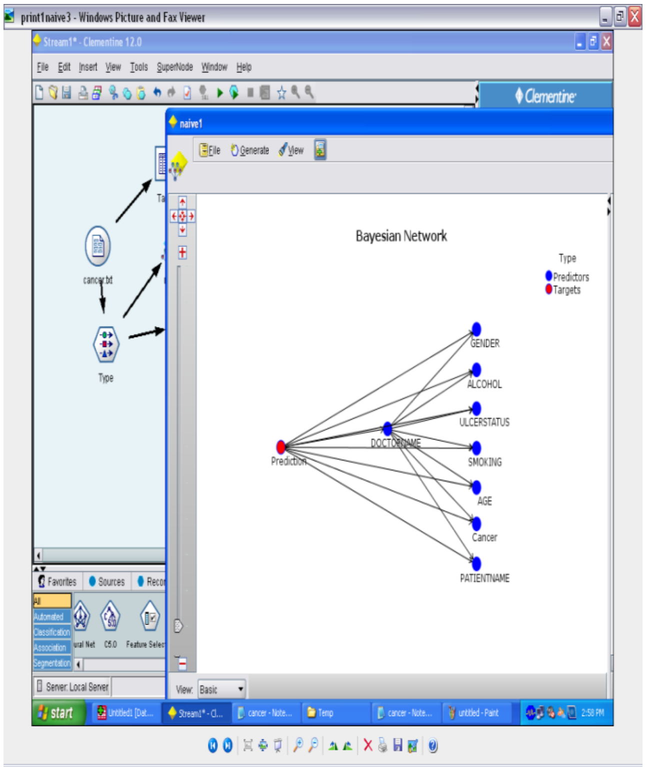
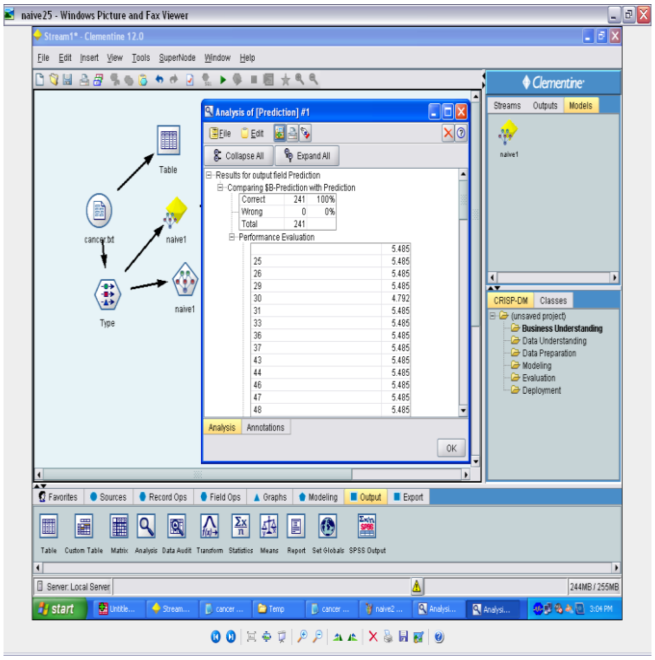


Figure5. Shows the working of Bayes Network.

The analysis of prediction in naïve bayes algorithm is shown below in Fig4.the analysis shows that for all the training tuples,the machine learning algorithm predicts the correct result. the results of the output prediction with the training tuples and testing tuples as shown in Fig6.


naive25 - Windows Picture and Fax Viewer
Stream1* - Clementine 12.0
File Edit Insert View Tools SuperNode Window Help
Clementine
Streams Outputs Models
naive1
Table
cancer.txt
naive1
Type
naive1
CRISP-DM Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export
Table Custom Table Matrix Analysis Data Audit Transform Statistics Means Report Set Globals SPSS Output
Server: Local Server
244MB / 255MB
start
3:04 PM

Analysis of [Prediction] #1

Results for output field Prediction

Comparing $B-Prediction with Prediction

Correct	241	100%
Wrong	0	0%
Total	241	

Performance Evaluation

	5.485
25	5.485
26	5.485
29	5.485
30	4.792
31	5.485
33	5.485
36	5.485
37	5.485
43	5.485
44	5.485
46	5.485
47	5.485
48	5.485

Figure6. Result in Bayes.

C. Implementation of SVM Algorithm:

The variable importance in support vector machine is shown in Fig7.

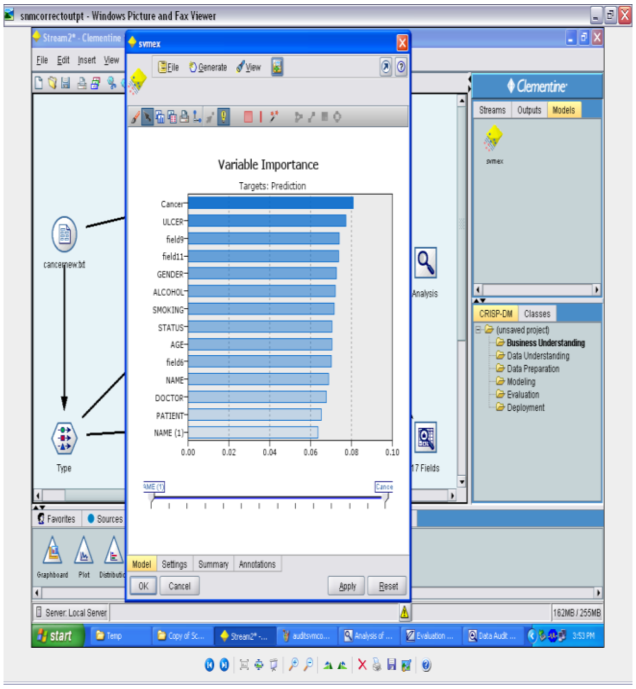
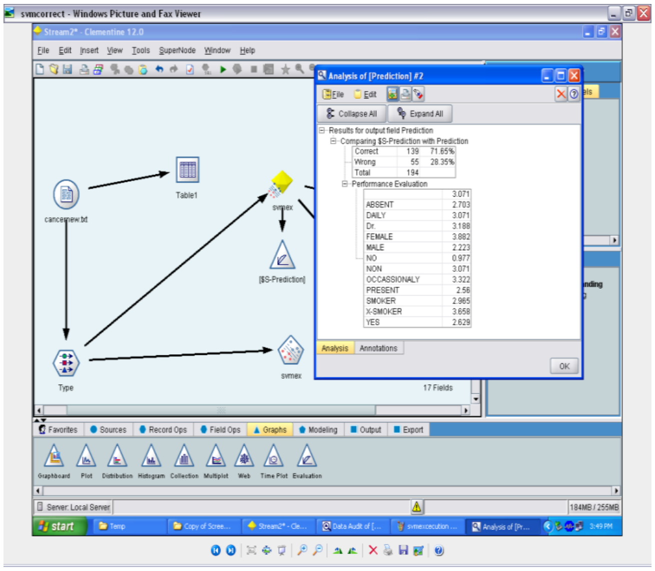


Figure7. Variable importance

In the SVM model we train the algorithm with the training data along with the training data the model is tested with the test data as shown in Fig7.in this model we have taken 139 training samples and 55 testing sample (i.e. wrong as shown in figure).



Results for output field Prediction

Comparing $S-Prediction with Prediction

Correct	139	71.65%
Wrong	55	28.35%
Total	194	

Performance Evaluation

	3.071
ABSENT	2.703
DAILY	3.071
Dr.	3.188
FEMALE	3.882
MALE	2.223
NO	0.977
NON	3.071
OCCASSIONALY	3.322
PRESENT	2.56
SMOKER	2.965
X-SMOKER	3.658
YES	2.629

Figure7: Result in SVM.

D. Comparison Of Naïve Bayes And SVM:

The naïve bayes algorithm is not able to handle the data subject to the noise, where as the support vector machine can handle the data in case of noise. the result of the naïve bayes method is shown in fig5.the method is trained with out noise, it shows the correct result and it is also tested with noise also and we are achieving only 48.45%.with the same noise data with tested the support vector machine method and we get 71.65% accuracy result.

VI. CONCLUSION

In this paper we implemented two machine learning algorithms for predicting oral cancer in the early stage by using some symptoms as specified in the paper and the accuracy of the result can be achieved by applying some other machine learning algorithms.

VII. REFERENCES

[1] Parvesh Kumar Siri Krishan Wasan "Analysis of cancer datasets using Classification Algorithms",IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.6, June 2010.

[2] http://www.peoples-health.com/oral_cancer.htm

[3] Multi spectral Optical Examination of Oral Lesions in High-Risk Populations By Mark Nichols , DDS ,Vice President of Clinical Affairs , Bering Omega Dental Clinic.

[4] Jiawei Han and Micheline Kambers,"Data Mining –Concepts and Techniques "2nd Edition, Morgan Kaufman Publications, 2005.

[5] Diana Dumitru ,"Prediction of recurrent events in breast cancer using the Naive Bayesian classification", Annals of University of Craiova, Math. Comp. Sci. Ser. Volume 36(2), 2009, Pages 92-96 ISSN: 1223-6934.

[6] R. Mallika, and V. Saravanan ,"An SVM based Classification Method for Cancer Data using Minimum Microarray Gene Expressions"World Academy of Science, Engineering and Technology 62010.

[7] Max welling ,"Support Vector machines",Department of Computer Science University of Toronto.

[8] Grace Bujewski, Brian Rutherford " The Rapid Optical Screening Tool (ROST) Laser- Induced Fluorescence (LIF) System for Screening of Petroleum Hydrocarbons in Subsurface Soils".