International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) ISSN: 2395-5325
Volume 3, Special Issue 3 (September '2017)
Proceeding of - International Conference on Recent Trends in Computational Life Science and Information Technology,
Conference held at Tiruppur Kumaran College for Women,Tirupur, Tamilnadu, India.

AN APPROACH TO ORAL CANCER PROGNOSIS USING GENE EXPRESSION DATASET

R.Vidhu,
Assistant professor,
Vidyasagar College of Arts and Science,
Udumalpet,Tamilnadu, India.

S.Kiruthika,
Assistant professor,
PKR Arts College for Women,
Gobichettipalayam,Tamilnadu, India.

C.B.Lakshmi,
Assistant professor
Izee College of Management and Information Science
Bangalore,Karnataka,India.

Abstract: Oral Squamous Cell Carcinoma (OSCC) constitutes the 8th most common neoplasm in humans. OSCC results from a combination of risk habit factors such as tobacco use, betel-quid chewing, alcohol consumption and genetic damage that leads to DNA alterations in key cellular genes. Diagnosis of oral cancer at its early stage will reduce the mortality rate. The diagnosis requires data collection from patients. Data mining techniques, such as pattern association, classification and clustering, are now frequently applied in cancer and gene expressions correlation studies. In this study, a framework for learning the structure of oral cancer genetic network based on DBNs is proposed. This approach improves the prediction level of the oral cancer and is tested with the gene data is set to prove the increase of classification accuracy and reduce execution time compared to the existing technique.

Keywords: Oral Squamous Cell Carcinoma, Association, Classification and Clustering gene expression data, DBNs.

I. INTRODUCTION

The tendency in recent decades to computerize the process of disease treatment ensures a more rapid accumulation of medical information. Information technologies are actively used in the sector of health protection. National electronic health records systems and medical imaging archives are implemented all over the world. Health care institutions implement and deploy hospital information systems (HIS), radiological picture reviewing and archiving systems (PACS), laboratory information systems (LIS), and others. Medical information systems (hereinafter – MIS) accumulate a structured medical history of a patient which includes classified attributes, such as diagnosis, patient demographic data, vital functions, test results, and unstructured data, such as images and video files. Analysis and mining of this data are strategically significant to the health sector and important to each patient. An intellectual analysis of the accumulated data offers new instruments for the following tasks: faster patient diagnosis, selection of optimal treatment, prediction of treatment duration and its outcome, determination of complication risks, and optimization of healthcare facility resources.

1.1 Data Mining in Healthcare and Medicine: overview and analysis

(DM) is in its infancy. Over the past decade, the application of DM in biomedicine has also been actively investigated. A noticeable increase in the number of publications and presentations at conferences indicates the relevance of this topic. Although it is not the first decade that methods of DM are being applied in medicine globally, practical application beyond research is still considered to be innovative and challenging.

1.2 Oral cancer -Detection

The indications for an oral cancer at an earlier stage are: 1) Patches inside the mouth or on lips that are white, red or mixture of white and red 2) Bleeding in the mouth 3) Difficulty or pain when swallowing 4) A lump in the neck. These indications should raise the suspicion of cancer and needs proper treatment. Therapies for Oral Cancer include surgery, radiation therapy and chemotherapy.

Some of the sections that are present in the minimum data set for mental health are the following:

- Name and identification numbers
- Referral items
- Mental health service history
- Assessment information
- Mental state indicators
- Substance use and extreme behavior
- Harm to self and others
- Behavior disturbance
- Self-care
- Medications
- Health conditions and possible medication side effects
- Service utilization and treatment

II. DATASET FORMULATION

In the sections that follow, we describe the input variablecomprising each source of data as well as the preprocessing steps employed to enhance the quality of input data.

1) Clinical Data: The specific features considered during this study from a clinical aspect are shown in Table I, and are in accordance with current medical domain knowledge. Certain features, however, contain missing values for a considerable

number of patients. Specifically, features with more than 90% of missing values (indicated with bold in Table I) are completely discarded from our analysis, whereas, the values of features with less percentage of missing values are imputed with modes and means in the case of nominal and numerical features, respectively.

2) Imaging Data: As for the imaging data, 17 features are extracted from CT and MRI images of the head and neck area. For the case of imaging data, none of the features contain missing values; therefore, no imputation procedure is needed and the feature vector is used as it is.

3) Genomic Data: From the cancerous tissue specimen of each patient, we extract the expression of 45 015 genes. All microarray experiments have been conducted with the same platform and feature extraction software in order to maintain solely biological variability and avoid all other potentially contaminating sources. The extracted gene expression files are initially subject to a series of basic preprocessing steps.

Table I-Clinical Dataset

Ecog Status	Mobile Prosthesis	BMI	Grade Of Differentiation
Weight	Dental Cusps	Substance Exposition	Surgical Margins
Height	Galvanic Current	Precancerous Lesions	Martinez-Gimeno Score Anneroths Mod Score
Diabetes	Oral Hygiene	**Duration**	
Allergies	Infection	Immunosuppressor Treatments Presence	D2_40Stain
Cholesterol	**Type Of Infection**	**Immuno Duration**	P53_STAIN
Hypertension	Physical Agents	**Immuno Type**	P16Ink4aStain
Family History Of Malignance	**Type Of Physical Agent**	Tumor Maximum Diameter	EGFR Stain
Smoker	Diet Deficit	Tumor Thickness	CyclinD1Stain
Smoking Habits	**Fe Haematic Concentration**	Depth Of Invasion	Ki67Stain
Quantity Per Day	Plummer Vinson	Basaloid Features	HPV_DNA
Smoking For	**Hb Haematic Concentration**	Lympho Plasmacytic Rection	T Staging
Ex Smoker	**B12 Vitamins Haematic Concentration**	Lympho Plasmacytic Invasion	N Staging
Quitted Smoking	**A Vitamins Haematic Concentration**	Perineural Invasion	M Staging
Alcohol	**E Vitamins Haematic Concentration**	Degree Of Cells Keratinisation	
Drinking Habits	**Folati**	Nuclear Pleomorphism	
Mechanical Trauma	Eating Habits	Number of Mitoses per 10HPF	

Note: Features in bold have >90% missing values.

2.2 OSCC Dataset

The patients who havebeen diagnosed with OSCC and had reached completeremission, genomic data from circulating blood cells had been collected, at the baseline state and during scheduled visits, in consecutive time intervals of each patient. Consequently, patients had been discriminated into two groups, namely relapsers and non-relapsers based on the occurrence or not of a disease relapse during the follow-up period. More specifically, 12 out of 23 patients had already suffered a recurrence, while the remaining 11 were still disease-free, during the follow-up period. Initially, the genomic data are applied to preprocessing steps in order to avoid any systematic variations. After the employment of an algorithm for microarray analysis, a significant number of differentially expressed genes was summarized and the quality of our dataset was enhanced. The retained genes (n=9) along with those that were extracted from the literature as oral cancer risk associated genes (n=28), were then fed as input to the next step of our methodology aiming to infer their interaction network.

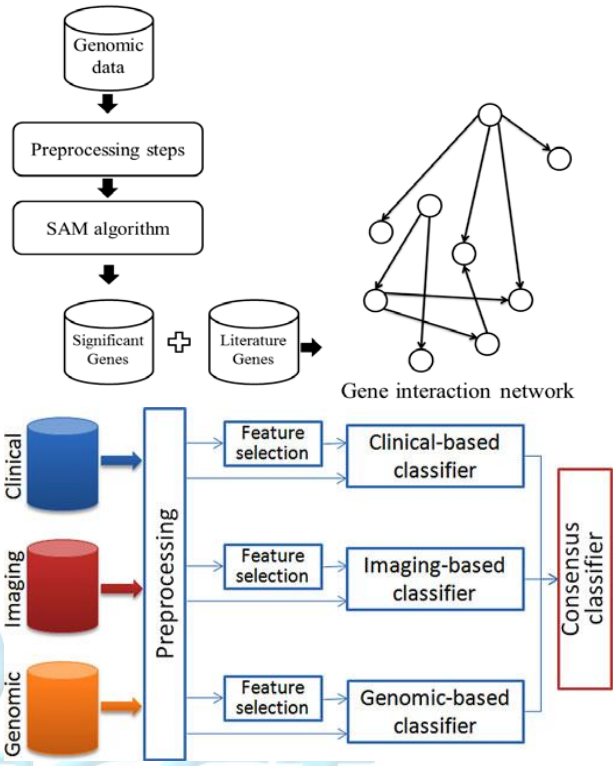

Figure 1:Gene dataset and processing method

III. DYNAMIC BAYESIAN NETWORKS

After the dataset of the 37 genes for each one of the 23 patients had been formulated, it was fed as input to the next step of our analysis in order to learn the structure of the gene interaction network. It should be clarified that the input file for our analysis contains the expression values of 37 genes in two time slices, i.e (i) the baseline (t) and (ii) the follow up (t+1). BNFinder2 was employed aiming to infer the structure of the networks from our experimental expression Data. DBNs can be considered as temporal extensions of Bayesian Networks (BNs) .They can be used for "exploring" biological networks in terms of temporal changes of nodes (genes and proteins) as well as of formation of new nodes or removal of existing, over timeslices.The employment of DBNs to formulate the causal relationships among variables can be defined as the process of inferring the possible interactions between genes from experimental genomic data and through computational analysis.

As mentioned above, we have constructed gene networks using DBNs, thus, producing directed acyclic graphs (DAGs). These networks were then fed as input to Cytoscape for visualization purposes, as well as, for further functional and topological analysis. Regarding the gene networks inferred separately for each group of patients,we subsequently mapped their interactions and extract the significant nodes in terms of topological analysis. Theintegration of gene expression data with network

knowledge allowed us for the discovery of important nodes related to
oral cancer recurrence. In the current report, time series geneexpression data are exploited in order to predict oral cancer recurrence using DBNs.

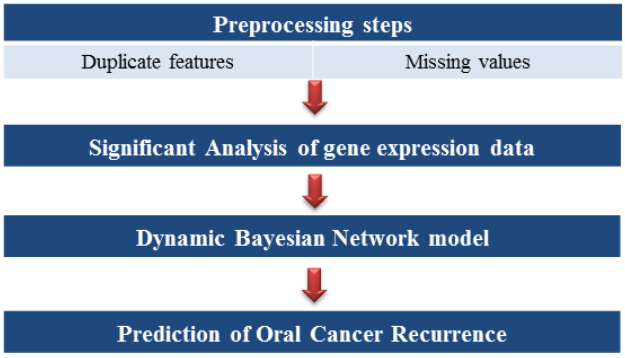

Figure2: Preprocessing steps in DBN

3.1 DBNs-An effective method to predict OSCC

The approach of Dynamic Bayesian Networks has been widely used for the inference of gene regulatory networks from time series microarray data; thus, they constitute an appealing choice for modeling oral cancer recurrence. DBNs are an extension of Bayesian Networks (BNs) which encode the joint probability distributions over a set of random variables X □ {x1, ...,xn}

DBNs are defined by a graphical structure and a set of parameters. Therefore, in order to construct a DBN we need to specify the intra-slice topology (connections within a slice), the inter-slice topology (connections between two slices) and the parameters for the first two slices. Figure 3 depicts a simple structure of a DBN with two time slices (t=0 and t=1). Intra and inter-slice topology can be observed between the variables. We employ the training data in order to define the structure and the parameters of two DBN models related to the status of a specific patient, i.e. relapser or norelapser. The parameters were specified among the variables within the first time slice and across the first and the second time slice. As DBNs capture temporal causalities between the state variables, they can therefore depict a better approximation of the actual stochastic process being studied.

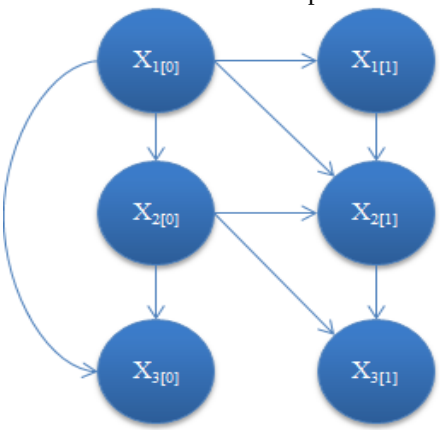

Figure3: A simple example of a DBN structure

3. 2 Benefits of DBN

Based on that knowledge our aim is to infer the corresponding dynamic Bayesian networks and subsequently conjecture about the causal relationships among genes within the same time-slice
and between consecutive time-slices. Moreover, the objectives of the proposed methodology are to: (i) accurately estimate the patient prognosis regarding oral cancer recurrence, and (ii) provide better insights into the underlying biological processes of the disease.

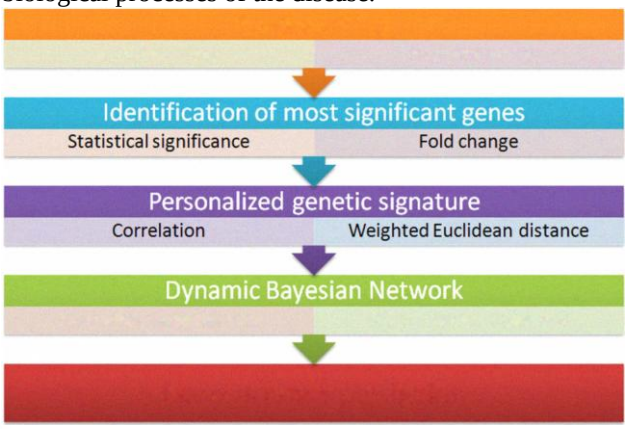

Figure4: Flowchart of the DBN methodology.

IV. CHALLENGES IN DATA MINING FOR HEALTHCARE

- Missing values, noise, and outliers
- "Cleaning data from noise and outliers and handlingmissing values, and then finding the right subset of data,prepares them for successful data mining".
- Transcription and manipulation of patient records oftenresult in a high volume of noise and a high portion of missing values.
- "Missing attribute values can impact the assessment ofwhether a particular combination of attribute-value pairs issignificant within a dataset"

V. CONCLUSION

Oral cancer being a deadly disease should be treated with care. Early diagnosis will surely reduce the death rate of patients. DBN will predict OSCC and will deliver the technology and knowledge that users need to readily: (1) organize relevant data, (2) detect cancer patterns (3) formulate models that explain the patterns, and(4) evaluate the efficacy of specified treatments and interventions with the formulations. Future work shall involve applying hybridization of data mining algorithms using the gene data set and identifying the useful patterns.

VI.REFERENCES

[1] L. J. van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, and A. T. Witteveen, "Gene expression profiling predicts clinical

outcome of breast cancer," *Nature*, vol. 415, pp. 530-536, 2002.

[2] G. Wu and L. Stein, "A network module-based method for identifying cancer prognostic signatures," *Genome Biol*, vol. 13, p. R112, 2012.

[3] M. A. Mutalib, L. E. Chai, C. K. Chong, Y. W. Choon, S. Deris, R. M. Illias, and M. S. Mohamad, "Inferring Gene Networks from Gene Expression Data Using Dynamic Bayesian Network with Different Scoring Metric Approaches," in *Advances inBiomedical Infrastructure 2013*, ed: Springer, 2013, pp. 77-86.

[4] Prediction of Oral Cancer Recurrence using Dynamic Bayesian Networks KonstantinaKourou, George Rigas, Konstantinos P. Exarchos, Costas Papaloukas and Dimitrios I.Fotiadis, Senior *Member, IEEE*

[5] Meta-Analyses of Microarray Datasets Identifies ANO1 and FADD as Prognostic Markers of Head and Neck Cancer

[6] Towards building a Dynamic Bayesian Network for monitoring oral cancer progression using time-course gene expression data. Konstantinos P. Exarchos, George Rigas, Y orgosGoletsis and Dimitrios I. F otiadis, Senior Member,IEEE.

[7] A Comparison on Performance of Data Mining Algorithms in Classification of Social Network Data

[8] Development and Application of Data Mining Methods in Medical Diagnostics and Healthcare Management .International Journal of Engineering Research & Technology (IJERT)Vol. 3 Issue 1, January – 2014ISSN: 2278-0181

[9] Data Mining Issues and Challenges in Healthcare DomianDr. C. Sunil Kumar1, Dr. A. Govardhan 2, B. Sunil Srinivas 3

[10] Degoulet, P. and Fieschi, M. Introduction to ClinicalInformatics, Springer, New York, 1997. [6] Fayyad,U., Haussler, D. and Stolorz, P. "Mining ScientificData", Communications of the ACM,(39:11),1996, pp. 55-60

[11] Goodall, C. R. "Data Mining of Massive Datasetsin Healthcare", Journal of Computational &Graphical Statistics, (8:3), 1999, pp. 620-635.

[12] Lavrac, N. "Selected Techniques for Data Miningin Medicine", Artificial Intelligence inMedicine Journal, (16:1), 1999, pp. 3-23.

[13] [Persson09] Persson, M.; Lavesson, N., "Identification of Surgery Indicators by Mining HospitalData: A Preliminary Study", 20th International Workshop on Database and Expert SystemsApplication, 2009. DEXA '09

[14] Shantakumar B. Patil and Dr. Y.S. Kumaraswamy [2009] Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research 2009; ISSN 1450-216X Vol.31 No.4 : 642-656

[15] DSVGK Kaladhar, B. Chandana and P. Bharath Kumar, "Predicting cancer survivability using Classification algorithms", International Journal of Research and Reviews in Computer Science (IJRRCS) Vol.2, No.2, pp 340 – 343, April 2011.

[16] Werning, John W (May 16, 2007). Oral cancer: diagnosis, management, and rehabilitation. p. 1.ISBN 978-1588903099.

[17] A New Feature Selection Method for Oral Cancer Using Data Mining Techniques Mrs. R. Vidhu1, Mrs. S. Kiruthika2 ,International Journal of Advanced Research in Computer and Communication EngineeringVol. 5, Issue 1, January 2016

[18] http://www.oralcancerfoundation.org/facts/index.htm

[19] Sharma, N. and Om, H. (2012) Framework for Early Detection and Prevention of Oral Cancer Using Data Mining. International Journal of Advances in Engineering & Technology, 4, 302-310.

[20] www.yourtotalhealth.ivillage.com

[21] www.oralcancerawareness.org