A Study on Future Trends of Data Mining for Prediction of Cancer

1Aditi Nautiyal, 2Amit Kumar Mishra
1 M.Tech Student, Department of Computer Science & Engineering, DIT University, Dehradun, Uttrakhand-
2 Assistant Professor, Department of Computer Science & Engineering, DIT University, Dehradun, Uttrakhand-

ABSTRACT
There are so many diseases which are affecting human lives and it is required to detect the disease in the initial stage which can help in reducing mortality in all over the world. These days we have technology that can easily process bulk of data all together and we can get results very quickly. Among all the dangerous diseases we have seen that Cancer is the most common among people and taking lives of millions of people globally. With the help of various Data Mining algorithms we are able to predict any disease in the wink of an eye.

This technology can help in medical field by predicting any disease in its early stage. Under Data Mining there are various approaches i.e. Association Rule Mining, Regression, Classification, Clustering and Anomaly Detection. Algorithms that are applied on the data sets are Support Vector Machine, Decision Tree, Artificial Neural Network, K-Nearest Neighbour etc. The process which data mining follows is known as knowledge discovery in database and it comprises five stages i.e. Selection, Processing, Transformation, data mining and Evaluation. The motive is to use any of the algorithms which can help the researchers and doctors to know about the disease in the starting stage so that the patient can get the best treatment and can get rid of the disease quickly.

KEYWORDS
Feature Selection, Gene Selection, Machine Learning Techniques, Support Vector Machine, Decision Tree, Naïve Bayes, Artificial Neural Network, Data Mining.

1. INTRODUCTION

Cancer is a disease which involves abnormal cell growth and can affect any part of human body. There are over more than 100 different known cancers [1]. The cancer which is common in females is Breast Cancer [2]. Around 55,699 females were diagnosed in 2003 and 39,805 died due to acute breast cancer[3][4]. To overcome this major issue of death due to cancer disease researchers have been using machine learning techniques in Biomedical or Bioinformatics for the last two decades to predict the disease beforehand[5]. Early diagnosis can prevent the danger as the patient can get treatment at right time. In Cancer prediction the scientists and researchers have following issues in their mind and the issues are - Cancer Susceptibility, Cancer Recurrence and Cancer Survivability. In susceptibility, probability of happening the disease is checked before the occurrence of it. In recurrence, the chances of recurrence of cancer among the patient is checked and the last one is cancer survivability in which the life expectancy of the patient after the treatment of cancer is analyzed.

These three things can be done with the help of Machine Learning (ML) algorithms. ML is a technique through which a machine can mimic human decision taking power and performs accordingly. The fundamental work of ML technique is to identify meaningful information from the bulk of information and predict the output or relevant results. These days ML techniques or ML algorithms are used in recommendation systems, healthcare, financial trading, online search and natural language processing etc. The types of ML techniques are discussed belowi) Supervised learning- In this solution are called target and situation are called input or unlabeled data. These two things are joined to form labeled data. Hence when we trained our machine in such a way that for every input we get a corresponding output.

It is divided into two categories “Regression” and “Classification”. The first one tries to find a function which can model the data with least error and the second one category the data into classes. Some supervised learning algorithms are - Support vector machine (SVM) for classification problem, Decision Tree(DT) , Random forest (RF) , K- Nearest Neighbor(KNN), Artificial Neural Network(ANN).

ii) Unsupervised Learning- In this we have input data but no corresponding output data. It is further categorized into "Clustering" and "Association". In clustering the data is represented in the forms of different groups or clusters. Association rule is when the methods or rules are discovered to describe the large portion of the data ,for example students X who read book Y also read book Z. The algorithms are K means and Apriori .

Researches have already been done in biomedical field using different type of data sets [6-9]. Mostly four types of data sets are used i.e. Proteomic data (Mass Spectral Analysis, 2D Gel Data ,Specific Protein Biomarkers), Clinical data (Tumor Staging, Age ,Weight, Tumor Size, Histology etc.) ,Genomic data (Deoxyribonucleic Acid (DNA) Sequencing, Microarray, SNP's ,Mutation) and Imaging(Functional Magnetic Resonance(fMRI), Positron Emission Tomography(PET) ,Micro-CT). Molecular Biomarkers and pattern of tumor protein have shown good predictive indicators [10][11].

The data sets which are mentioned above can be a huge collection of patterns and can have combination of both relevant and irrelevant data. Now to reduce the number of attributes of lower ranked data and to select the relevant data from the huge data set Feature Selection
2
Technique has to be used that can improve the accuracy of the algorithm. It not only reduces the dimensionality but also increases the running speed of the algorithm. Methods of feature selection are- Filter methods, Wrapper methods and Embedded scheme[12].
The next point is how the performances of different algorithms can be measured. The evaluation can be done by using any of the algorithm discussed below-

Random sampling , it selects the training set and testing set randomly and repeats the holdout method several times. Algorithms are selection rejection algorithm, naïve algorithm etc. Cross Validation tests the model in the training phase to avoid overfitting. Types of cross validation are- Exhaustive and Non Exhaustive cross validation. Last approach is Bootstrap in which test relies on random sampling with replacement. Types of bootstrap are- Bayesian bootstrap, smooth bootstrap, block bootstrap etc.

2. ML TECHNIQUES

Earlier cancer patients had to undergone painful and slow diagnostics clinical courses but due to the continuous research in the field of machine learning and testing techniques we are able to find patterns from the bulk of data set or information about the past history of the patients report. ML algorithms are discussed below which are used widely in prediction of cancer and detecting it in early stage.

2.1 SUPPORT VECTOR MACHINE

It comes under predictive model and used for classification purpose. It draws either a single hyperplane or multiple hyperplane and if the margin is larger then the error is minimum. In an n-dimensional space, the hyperplane is of (n-1) dimension with flat subspace that need not pass through the origin. If there does not exist linearly separable hyperplane for dataset, linear classifier can't be formed in that case.
Kernel trick have to be applied to maximum-margin hyperplanes to develop non linear classifier. Non linear kernel function will be applied to the hyperplane in replacement of dot product. Cubic ,quadratic or higher-order polynomial function, Sigmoid function and Gaussian radial basis function are forms of non linear kernel function. Results have shown that it gives good results in predicting cancer[13].

2.2 ARTIFICIAL NEURAL NETWORK

The idea of Artificial Neural Network is derived from biological neural system. It is the interconnection of large number of units or nodes which can establish a good communication with one another. The units are also called neurons and they all operate simultaneously
as biological neural system. Neurons are joined together through connecting links and all connecting links have knowledge about the inputs. This system is fault tolerant, gives robust performance and was used a lot in biomedical field earlier [14]. The diagram of ANN is
shown below in fig 2
3

2.3 DECISION TREE

It enables the researchers to come to a right solution. It gives a structure as of a tree and the branches of the tree are called as nodes. The
uncertainty nodes are represented by circles, the decision are represented by squares and the end nodes are represented by triangles. It is

used in several fields like where we have to reach to a particular aim. Prediction of various cancer diseases can be done using this technique. The figure of decision tree is shown below in figure 3.
Fig2. Decision Tree

2.4 K-NEAREST NEIGHBORS
It performs great in pattern recognition and predictive analysis. For any new data point, K-NN gather data points that are close to it. Any attributes that can vary on a large scale may have effective impact on the distance between data points [15]. The algorithm sort these closest data points in terms of distance from the arrival data point. This distance can be measure in various way but Euclidian distance is the suggested one by experts. Next step is to take a specific number of data points whose distance are lesser among all and then
categorize those data point. The category with highest number of data point will be the category of the new data point.

2.5 RANDOM FOREST
It belongs to predictive model that contains classification and regression methods. Forest means collection of tress , therefore the name means group of trees together in a common place .It gives many classification trees without pruning. Each classification tree gives a certain number of votes for each class. Among all the trees, the algorithm chooses the classification with the most number of votes. Random forest runs efficiently on large datasets but is comparatively slower than other algorithms. It can effectively estimate missing values and hence is suitable for handling datasets with large number of missing values.

3. RELATED WORK USING DATA MINING
Studies have shown that already a lot of research works have been done on disease prognosis using various data mining techniques yet scientists are trying to develop a new technique from the already existing techniques to achieve more accuracy. Various scientists have used decision tree algorithm to predict cancer recurrence[16,17].It is necessary to efficiently recognise microarray gene expressions and that can be done using DNA microarray technology. Fuhrman, Stefanie et al. have applied Information Theory to gene identification problem[18]. Friedman, Nir et al have proposed Bayesian Network and Arkin, Adam et et al. have used reverse engineering
method[19][20]. In classifying gene expression data various machine learning techniques have been used earlier. Dudoit proposed Fisher
linear discriminant analysis, Li, Leping et al. have proposed K-Nearest Neighbour[21]. Khan, Javed et al. have proposed Decision Tree and Xu , Yan et al. have proposed multi-layer perceptron[22][23].Furey, Terrance et al have proposed Support Vector Machine, Brown,Michael et al proposed boosting[24][25]. Jayashree, Dev et al. have focused on three different classification techniques and concluded good result[26]. Breast cancer is another risky disease that causes death for numerous ladies all over the world. The classifier obtained by supervised machine learning techniques will be very supportive in the field of medical disorders and proper diagnosis.

Bellaachia and Guven used various mining techniques and limited attributes to conclude that among all the used techniques decision tree gave the comparatively best output [27]. Jahanvi, Joshi et al. compared two techniques i.e. KNN and Expectation Maximisation (EM) and found that EM algorithm can also give good results as KNN[28]. Machhale, K et al. proposed a highly distribution framework to observe typical and anomalous Magnetic Resonance Image (MRI) brain images[29]. Gayathri, Sumathi et al. have done assessment on various algorithms like Relevance Vector Machine, Support Vector Machine, Aritificial Neural Network[30].

4. CONCLUSION
In this paper details of various machine learning techniques which are used in predicting cancer disease as well as cancer recurrence and susceptibility are mentioned. Feature selection algorithm can remove unwanted and unimportant data for more accurate results. Various type of data has been used like gene expressions, micro arrays, DNA samples etc to predict cancer in early stage. Two or more model can be hybrid for more better results. In most of the cases we have seen that SVM gives good results but in some other cases we have also seen that the other machine learning techniques give even more better results. This paper can be helpful for researchers to know about the various ML techniques and the results of the algorithms in finding cancer diseases in early stage.

5. ACKNOWLEDGMENTS
I would like to give gratitude to Dr. Arvind Kumar Tiwari, Associate Professor KNIT,Sultanpur who has motivated me to study about this research topic and my special thanks to all the researchers who have written the papers which I thoroughly studied and got more idea about the ML algorithms.

6. REFERENCES

[1]Shandilya, S. and Chandankhede, C. Survey on Recent Cancer Classification Systems for Cancer Diagnosis Department of Information Technology MIT,Pune.

[2] Breast cancer statistics. [Online]. Available:http://www.wcrf.org/int/cancer-facts-figures/data-specific-ancers/breastcancer-statistics, accessed on: Aug. 25, 2017.

[3] Calle, J. 2003-2004. Breast cancer facts and figures.

[4]American Cancer Society 2004 , Breast cancer Q & A facts and statistics, BCI.

[5]Hanahan, D. and Weinberg R.A. 2011.Hallmarks of cancer: The next generation. Cell ,144:646–74.

[6] Fortunato, O. , Boeri, M., Verri, C., Conte, D. , Mensah, M. , Suatoni, P. 2014 Assessment of circulating microRNAs in plasma of lung cancer patients. Molecules , 19:3038–54.

[7] Heneghan, HM. , Miller, N. , Kerin, MJ. 2010 MicroRNAs as Biomarkers and Therapeutic targets in cancer, 10: 543–50.

[8] Madhavan, D. , Cuk, K. , Burwinkel, B. , Yang, R. 2013, Cancer diagnosis and prognosis decoded by blood-based circulating microRNA signatures. Front Genetics ,4.

[9] Zen, K. , Zhang, CY. 2012 CirculatingmicroRNAs: A novel class of biomarkers to diagnose and monitor human cancers. ,32:326–48.

[10]Piccart, M. , Lohrisch, C. , Di Leo, A. 2001 The predictive value of HER2 in breast cancer. Oncology, 61(Suppl 2):73–82.
5
[11] Savage, KJ. , Gascoyne, RD. 2004 Molecular signatures of lymphoma. Int J Hematol, 80:401–9.

[12]Ladha, L. and Deepa, T. Feature selection methods and Department of Computer Science Coimbatore, Tamilnadu, India.

[13] Pritom, A.H. and Munshi A.R. 2013 Predicting Breast Cancer Recurrence using Effective Classification and Feature Selection technique.

[14]. Simes, RJ. 1985 Treatment selection for cancer patients: application of statistical decision theory to the treatment of advanced ovarian cancer. J Chronic Disease ,38:171–86.

[15] James, G, and Witten, D. and Hastie, T. And Tibshirani, R. 2013 An Introduction to Statistical Learning, 1st edition.

[16] Zhou, ZH. and Jiang, Y. 2003 Medical diagnosis with C4.5 Rule preceded by artificial neural network ensemble. Biomedical 7: 37-42.

[17] Delen, D. Walker, G. and Kadam, A. 2005. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine 34: 113-127

[18] Fuhrman, Stefanie, et al. 2000 The application of Shannon entropy in the identification of putative drug targets . Biosystems 55.1 :5-14.

[19] Friedman, Nir, et al. 2000 Using Bayesian networks to analyze expression data. Journal of computational biology 7.3-4 : 601-620.

[20] Arkin, Adam, Peidong Shen, and Ross, J. 1997. A test case of correlation metric construction of a reaction pathway from measurements. Science 277.5330 : 1275-1279

[21] Li, Leping, et al. 2001. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of

parameters of the GA/KNN method , Bioinformatics 17.12 :1131-1142.

[22] Khan, Javed, et al. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature medicine 7.6 : 673-679.

[23] Xu, Yan, et al. 2002. Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer. Cancer Research 62.12: 3493-3497.
[24] Furey, Terrence, S. et al. 2000. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16.10 : 906-914.

[25]Brown, Michael, PS. , et al. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines.
Proceedings of the National Academy of Sciences 97.1 :262-267.

[26] Dev, Jayashree, et al. 2012. A Classification Technique for Microarray Gene Expression Data using PSO-FLANN. International Journal on Computer Science and Engineering 4.9 : 1534.

[27]Bellaachia, Abdelghani, and Erhan, Guven. 2006. Predicting breast cancer survivability using data mining techniques. page 58.13 :10-110.

[28]Joshi, Jahanvi, Doshi, R. and Patel, J. 2014. Diagnosis of breast cancer using clustering data mining approach. International Journal of Computer Applications 101.10.

[29] Machhale, K. , Nandpuru H.B. , Kapur, V. and Kosta, L. 2015. MRI brain cancer classification using hybrid classifier (SVMKNN), in International Conference on Industrial Instrumentation and Control(ICIC), Pune, pp. 60–65.

[30] Gayathri, B.M. , Sumathi, C.P. and Santhanam, T. 2013 Breast cancer diagnosis using machine learning algorithm- A survey, in International Journal of Distributed and Parallel System(IJDPS) Vol.4, No.