Literature Review on Data Mining Technique Used For Diagnosis and Prognosis of Cancer Disease

[1]Samiksha Zaveri, [2]Dr.Kamini Solanki
[1]Ph.D Scholar, [2]Assistant Professor
[1]Computer Science and Application
[1]Parul University, Vadodara, India

Abstract: Cancer is one of the leading and fatal diseases in the world. This paper represents the overview, advantages and disadvantages of the current research being carried out using different data mining techniques for the diagnosis and prognosis of cancer diseases. It also discusses the survey of cancer disease over the past decade that helps the researchers for further research for better result.

IndexTerms - **Data Mining,Machine learning algorithms,Cancer disease Diagnosis,Prognosis.**

I. INTRODUCTION

Data mining is "a process of nontrivial extraction of implicit, formerly unknown and useful information from the data stored in a database". Medical databases have a vast amount of data. Specific utilization of an automated system required to study of variety of methods available because there is adequacy of powerful mechanism to determine unknown information. Here, in this paper we take an outline of the current research and highlighting vital problems exists in current development.

II. TYPE STYLE AND FONTS

There are two different approaches used for special purposes. Data mining is the process of discover patterns in large amounts of data to extract useful information from patterns. There are various techniques such as artificial intelligence, machine learning, neural network and statistics to control the power of the pattern recognition with different end. With the help of machine learning process we can develop artificial intelligence. These allow the machine to learn from the analyzed data or with experience. An ability to induce new knowledge from experiences that need intelligence. Thus, a large area within AI is machine learning.

2.1 Types of Machine learning algorithms

It is divided into following types:

2.1.1 Supervised Machine Learning

These types of problems can be grouped into classification and regression problems.
Classification: It assigns data into discrete categories either pass or fails (1 or 0).
Regression: It is used to predict a numeric or continuous value (Percentage).

2.1.2 Unsupervised Machine Learning

It can be further grouped into association and clustering problems.
Association: When you want to discover rules that illustrate large piece of data.
Clustering: When you want to discover the natural groups in the data.

2.1.3 Semi-supervised Machine Learning

When you have a big amount of input data and only some of the data is labeled are known as semi-supervised learning.

III. OVERVIEW, ADVANTAGES AND DISADVANTAGES OF MACHINE LEARNING METHODS

Table 3.1: Method Name, Meaning, Advantages and Disadvantages

Method Name	Overview	Advantages	Disadvantages
Decision Tree	It is used for categorical and continuous dependent variables. This is done based on significant attributes and the occurrence of one does not affect the probability of other variables to make groups.	Feature selection is done automatically. Not required any efforts for data preparation . Not sensitive to outliers.	If new training set gets added ,It needs to be remodeled. Prone to over fitting especially with a lot of features. A solution is to stop the growth of the tree before it reaches one data point. It fits to noise.
SVM(Support Vector Machine)	It can be design each data set as a point in an n-element space (where n is the number of individual measurable property you have), with the value of each feature being the value of a	It works well in complicated domains, with outliers and for non-linear classifications.	Doesn't work well for large data sets and with a noise. The selection of a Kemel might be difficult.

	particular coordinate.		
Naive Baves	It is probabilistic classifier applying on Baves theorem with strong independence among predictors. It is assumes that the presence of a particular feature in a class is not related to the being there of any other feature.	Manages large data sets and with a lot of noise. Performs well in multi-class predictions.	It assumes that all properties separately give to the probability almost impossible in real-life. If a categorical variable has a category in the test dataset which was not in the training data, then the category gets assigned zero probability and the model is unable to predict.
KNN (K-Nearest Neighbours)	It is easy to repository all possible cases and arrange in the classes or case by a majority vote of its K Neighbours. The case ascribe to the class is the most common amongst its K-nearest neighbours, measured by a distance function.	Robust to noisy training data. It can detect linear or non – linear distributed data. It is non-parametric; therefore it takes no assumptions about the underlying data or its distribution.	It is non-parametric, it is actually slow in querying. It can be sensitive to outliers.
Random Forest	It is a collection of decision trees known as a forest. To classify a new object based on attributes, each tree gives a classification and we say the tree "votes" for that class. The forest chooses the classification with the most votes.	It handles very well high-dimensional spaces & reduces high variance. It does not expect linear features or features that interact linearly. It improves the robustness.	Lack of interpretation. It somehow violates the OCKHAM's RAZOR LAW, which implies that simplicity leads to greater accuracy.
Logistic Regression	It used for binary classification problems, meaning those in which there are two possible outcomes that are influenced by one or more explanatory variables. The algorithm estimates the probability of an outcome given a set of observed variables.	Robust to noise. Output can be interpreted as probability.	Hardly handle categorical (binary) features. Handles only linear decision boundaries, unless we create more features into polynomial terms.
Linear Regression	It used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.		It is limited to linear relationship. It is sensitive to outliers. Data must be independent.
Back Propagation	It leam classification model by training a multilayer feed-forward neural network with one input layer, some hidden layers, and one output layer. Each layer contains some units or perceptron each unit might be linked to other by weighted connections. The values of the weights are initialized before the training. The number of units in each layer, number of hidden layers, and the connections will be empirically defined at the very start.	It improves the accuracy of predictions in data mining and machine learning.	
Association Rules	An association rule is a rule which associate on relationships among a set of objects in a database. Given a set of transactions, where each transaction is a set of literals, an association rule is an expression of the form X Y where X and Y are sets of items. The transaction of the database which contain X tend to contain Y.		
Apriori algorithm	It uses a generate-and-test approach generates candidate item set and tests if	Easy execution. It uses apriori property for	It explains only the presence or absence of an item in the

	they are frequent I generation of candidate item sets is expensive (in both space and time) I support counting is expensive subset checking (computationally expensive) I multiple database scans (I/O)	pruning therefore; item sets left for further support checking remain less.	database. It scans the complete database multiple times. It has a complex candidate generation process that uses most of the time, space and memory. It works well only for small databases with large support factor.
FP-Growth	Two step approach: Step 1: Build a compact data structure called the FO-tree built using 2 passes over the data-set. Step 2: Extracts frequent item sets directly from the FP-tree traversal through FP-tree.	It scales than Apriori algorithm. It requires only two scans of the database without any candidate generation. It is not influenced by support factor.	The result is not unique for the same logical database. It cannot be used in interactive mining system. It cannot be used for incremental mining.
Clustering	Clustering is the task of segmenting a diverse grout into a number of similar subgroups or clusters. In clustering there are no predefined classes. The records are grouped together on the basis of self similarity.		
Partitioning	Suppose we are given a database of 'n' objects and the partitioning method constructs 'k' partition of data. Each partition will represent a cluster and k < n. It means that it will classify the data into k groups, which satisfy the following requirements. Each group contains at least one object. Each object must belong to exactly one group.	Relatively scalable and simple. Suitable for data sets with compact spherical clusters that are well-separated.	Degradation in high dimensional spaces. Poor cluster descriptors High sensitivity to initialization phase, noise and outliers.
Hierarchical	This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed.	Embedded flexibility regarding the level of granularity. Well suited for problems involving point linkages, e.g. taxonomy trees. Application to any attributes types.	Inability to make correction once the splitting decision is made. Lack of interpretability. Vagueness of termination criterion. Prohibitively expensive for high dimensional and massive data sets.
Density Based	This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold.	Discovery of arbitrary-shaped clusters with varying size. Resistance to noise and outliers.	High sensitivity to the setting of input parameters. Poor cluster descriptors. Unsuitable for high-dimensional datasets.

IV. REVIEWS OF RELATED RESEARCH WORK

A Priyanga S Prakasam (2013) gives the earlier warning in a cost effective manner to the users, navie bayes and decision tree are found to be best predictor and prominent method utilized for breast cancer prognosis and diagnosis in medical predication. ID3 algorithm provides highest accuracy in skin and lung cancer predication. A. Deivendran, Ms K. Yemuna Rane. (2014) developed a new hybrid intelligent technique based on association rule mining (ARM) and Neural Network (NN) which uses an Evolutionary Algorithm (EA) to deal with dimensionality problem for the diagnosis of breast cancer. GERM was used to ensure the finest selection of the input vector, while NN used for classification. A. Soltani Sarvestani, A. A. Safayi, N.M. Parandeh M. Salehi. (2010) used neural network with Feed forward Back propagation algorithm for breast cancer tumer. They classify the tumor from a symptom that caused the breast cancer which has hidden layer to achieve the highest accuracy among others. This is because; the mean square error is small and towards 0.001. Abdelghani Bellaachia Erhan Guven, Alaa M. El-Halees, Asem H. Shurrab. (2017) examined classification and clustering to predict and identify probable cancer patients. Therefore a novel multi layered method combining clustering and decision tree technique is used to build a cancer hazard predication system. Aicha Boutorh Ahmed Guessoum proposed system to predict lung, breast, oral, cervix, stomach and blood cancers and it is user friendly and cost saving too. Alaa M. El-Halees, Asem H. Shurrab. (2017), Bharathi, TS Arulananth, (2017)used k- means clustering algorithm to separate cancer and non cancer patient data. Additional, the cancer cluster is subdivided into six clusters. G. Ravi

Kumar, Dr. G. A. Ramachandra K. Nagamani, (2013) demonstrated association rules, rule induction and deep learning. Deep learning classifiers have the best ability to predict tumor types of blood diseases with an accuracy of 79.45%. On the other hand, rule induction and association rule gave satisfactory performance. Nakte Varun Himmatramka (2016) evaluated Several neural network structures such as statically neural network structure, self organizing map (SOM), radial basis function network (RBF), general regression neural network (GRNN) and probabilistic neural network (PNN). The performance of neural network structure was investigated for breast cancer diagnosis problems. RBF and PNN were proved as the best classifiers. However, the PNN gives best classification accuracy when the test set is considered. K. Arutchelvan, Dr. R. Periasamy. (Novermber 2015) suggested Decision Tree results are easier to read and interpret. K Arutchelvan Dr R Periyasamy, (2015) found that C4.5 algorithm has a much better performance than the other techniques. The prognosis problems is mainly analyzed under ANNs and its accuracy came higher in comparison to other classification technique applied for the same. Ramachandran N. Girija T. Bhuvaneswar, (2014) studied data classification algorithms, discrete particle swarm optimization (DPSO), a technique based on standard PSO has proved to be competitive in predicting breast cancer. They also confirmed that DPSO combined the proposed rule pruning is effective in predicting common types of cancer. Pallavi Mirajkar, G. Prasanna Lakshmi, (2017) introduced Self Organizing map (SOM) structure which is used to discover the hidden pattern in the lung disorder CT images by using the data ming techniques. This approach starts by extracting the lung regions from the CT image using image processing techniques. They analyzed different approaches to clustering of the SOM are considered.

Parshva Jain et al. (2017) studied Decision tree as classifier for breast cancer diagnosis. The classification accuracy depends on the exact metrics which are used to indicates the variety of features has been utilized. Seyed Mohammad Jafar Jalali et al. (2017) present a data mining framework for detecting breast cancer. They includes a feature selection procedure based on association rules, apriori algorithm. Dr. C Nalini et al. (2018) applied two prediction models for breast cancer Naive Bayes and J48.They classify four types of kidney diseases. Comparison of J48 and Naïve Bayes classification algorithms is done based on the performance factors classification accuracy and execution time. From the results, the Naïve bayesthat are accurate, hence it is considered as best classifier with minimum execution time.

R. Rajbharath et al. (2017) proposed a hybrid of Random Forest and Logistic Regression algorithms for building a breast cancer survivability prediction model. The Random Forest Technique is used to perform a preliminary screening of variables and to receive important ranks. Analysis results reveals higher accuracy and relatively a simple model. From the result of the analysis they indicate that the combination method of RF and LR is suitable for disease survivability prediction. The method has not only good classification accuracy but will result in relatively simple and interpretable model. Pallavi Mirajkar et al. (2018) integrated system which is based on combination of various data mining techniques such as analytical hierarchy process, rule based association, classification etc. that is helpful to predict the patient's disease status.

A.Daisy et al. (2017) reveals analysis of various data mining techniques used in classifying the cancer disease and to improve the accuracy of predicting the cancer in early stage and reduces the death rate. In this survey several data mining techniques such as Artificial Neural Network (ANN), Ensemble gene selection methods, pattern recognition, Learning Hidden Markov Models, random projection (RP), Ensemble Method, SVM classifier, Random Topology, Novel Gauss Newton Representation, Machine Learning, Decision tree, Sequential Minimal Optimization, Multiple filter multiple wrapper approach and Skewed gene selection algorithm etc used in the literatures and these methods have both merits and demerits. Initially the information gain is used for selects the significant features from the input patterns. Then the selected features are reduced by using the genetic algorithm (GA). Finally the gene expression profiles are utilizing to classify the human cancer disease chosen to improve the prediction of cancer classification. H Bharathi et al. (2017) developed a software based Self Organizing Map (SOM) structure which is used to discover the hidden patterns in the lung disorder CT images by using the data mining techniques. It starts with visualizing the closed structure of the Lung regions and then the disorder dataset is process using SOM Toolbox to create a learned SOM using K-means clustering. By using this data mining technique with SOM, it has the advantages of robust to analysis and cost effective method. The system extracts hidden knowledge from a historical lung cancer disease database.

Nimna Jeewandara et al. (2017) they were published by addressing Diabetes Mellitus, Heart Disease, Hypertension, Cancer and Hyperlipidaemia were used for this purpose. For the purpose of analysing the multiple diseases the researchers have used multiple stage analysis with many algorithms. As the non-communicable disease burden is a challenge many programs and research are conducted to prevent and control of these diseases all above researches used data mining approaches to predict the non-communicable diseases and the states by using many demographical, physiological, behavioral factors. They have used one or many data mining algorithms for one or more diseases to understand the hidden pattern and relationships of these diseases. They have been produced different accuracy, sensitivity and specificity according to their algorithms and the variables in large complex data sets. Chih-Jen Tsenga et al.(2017) determine the risk factors for women with ovarian cancer with regard to recurrence. they present a clinical data analysis that demonstrates that their methodology for identifying the potential risk factors has a much better predicting performance than that without the diagnostic scheme.

Rozita Jamili Oskouei et al.(2017) provide a comprehensive survey about applications of data mining techniques in breast cancer diagnosis, treatment & prognosis till now. They reviewed several research works which are done for diagnosis, treatment or prognosis breast cancers. Based on the results of this study, most of the research works are concerned on comparing the accuracy rate of data mining various algorithms. Shakuntala Jatav et al. (2018) has analyzed prediction systems for Diabetes, Kidney and Liver disease using Support Vector Machine(SVM) and Random Forest (RF). The performance of these techniques is

compared, based on precision, recall, accuracy, f_measure as well as time. They also show the study of different approaches such as neural network, naïve bayes, SVM, KNN, FCN, etc and it is concluded that SVM gives the best performance as compared to the other existing techniques.

V. PROPOSED WORK

In real world database there are noisy, missing and inconsistent data due to their massive size. The testing data are applied over Decision Tree and naive Bayesian (Hybrid) classification algorithm with feature extraction from data mining techniques to find the effectiveness of cancer prediction system. In decision tree, rules are extracted from the training dataset to form a tree structure, and this rule will be applied to the classification of testing data. The Navie Bayesian Classifier makes the assumption of class conditional independence .Use of a feature selection technique with a decision tree classifier to classify cancer. The feature selection method is helpful to analysis the features which have the high and low contribution that can impact to patient recovery factor. The classification method is a method which it tries to build decision making, and to support the doctor work. But, the precision of prediction result must be improved, because it is related with human life.

Algorithm for Decision Tree

Step 1: Collect Data set.
Step 2:Pre-processed data for Decision Tree-J48.(Remove noisy data and missing value).
Step 3: Generate a Decision Tree with leaf node
Step 4: Diagnosis of new paient is achived by new attributes value in Decision Tree and follow the path fill the leaf node reached.

Algorithm for Naive Bayesian

Step 1: Take training data set.
Step 2: Perform Pre- processing (Replace Missing Value).
Step 3: Calculate the probability of each attribute value.
Step 4: Apply Formula

$$P\ \left(attributevalue(ai)\ /\ subjectvalue\ (vj)\right) = n, c + mp\ /\ (n + m) \qquad (1)$$

Where: n = the number of training examples for which v = vj
nc = number of examples for which v = vj and a = ai
p = a priori estimate for P(aijvj)
m = the equivalent sample size
Step 5: Compare the value and classify the attribute value to one of the pre defined set of class.

VI. CONCLUSIONS

This paper provides a quick review of the different techniques in data mining. It is very difficult to name a single data mining algorithm as the best for the diagnosis and prognosis of all diseases. Depending on concrete situations, sometimes some algorithms perform better than others, but there are cases when a combination of the best properties of some of the aforementioned algorithms results more effective. The follow-up our work will aim at dealing with algorithms that have wider spectra of application for cancer diseases.

VII. ACKNOWLEDGMENT

I would like to thank my guide Dr. Kamini Solaki madam for encouraging me to write the paper and for their support and help.

REFERENCES

[1] A.Daisy , R.Porkodi. September- October 2017.A Survey On Cancer Classification Using Data Mining Techniques, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 6, Issue 5, ISSN 2278-6856.
[2] A Priyanga S Prakasam (2013). Effectiveness of data Mining – based Cancer Prediction System (DMBCPS), International Journal of Computer Applications (0975 – 8887) Volume 83 – No 10.
[3] A Priyanga S Prakasam. (2013). The Role of Data Mining-Based Cancer Prediction system (DMBCPS) in Cancer Awareness, International Journal of Computer Science and Engineering Communications- IJCSEC. Vol. 1.
[4] A. Deivendran, Ms K. Yemuna Rane. 2014. A Literature Review of Prediction Cancer Disease Using Modified ID3 Algorithm, International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue. ISSN 2348 -7968.
[5] A. Soltani Sarvestani, A. A. Safayi, N.M. Parandeh M. Salehi. 2010. Predication Breast Cancer Survivability Using Data Mining Techniques, 2nd International Conference on Software Technology and Engineering (ICSTE).
[6] Abdelghani Bellaachia Erhan Guven, Predicting Breast Cancer Survivability Using Data Mining Techniques.
[7] Aicha Boutorh Ahmed Guessoum, Classication of SNPs for Breast Cancer Diagnosis using Neural-Network-based Association Rules.
[8] Alaa M. El-Halees, Asem H. Shurrab. 2017. Blood Tumor Cancer Prediction Using Data Mining Techniques Health Informatics – An International Journal (HIIJ) Vol. 6, No.2.

[9] Chih-Jen Tsenga, Chi-Jie Lub, Chi-Chang Changc, Gin-Den Chena, Chalong Cheewakriangkrai d.2017. Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence, elesviwer-Artificial Intelligence in Medicine 78 47–54.

[10] Dr. C Nalini, D.Meera.2018. Breast cancer prediction system using Data mining methods, International Journal of Pure and Applied Mathematics, Volume 119 No. 12, 10901-10911.

[11] G. Ravi Kumar, Dr. G. A. Ramachandra K. Nagamani. 2013. An Effective Prediction of Breast Cancer Data using Data Mining Techniques, International Journal of Innovations in Engineering and Technology (IJIET) Vol. 2 Issue ISSN: 2319-1058.

[12] H Bharathi, TS Arulananth. 2017. A Review of Lung Cancer Prediction System Using Data Mining Techniques and Self Organizing Map (SOM), International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12 pp. 2190-2195.

[13] H Bharathi,TS Arulananth.2017. A Review of Lung cancer Prediction System using Data Mining Techniques and Self Organizing Map (SOM), International Journal of Applied Engineering Research, ISSN 0973-4562 Volume 12, Number 10 pp. 2190-2195.

[14] Jyotsna Nakte Varun Himmatramka. 2016. Breast Cancer Prediction System Using Data Mining Techniques, International Journal on Recent and Innovation Trends in Computing and Communicatio Volume: 4 Issue: 11 ISSN: 2321-8169.

[15] K. Arutchelvan, Dr. R. Periasamy. (Novermber 2015) Analysis of Cancer Detection System Using Data Mining approach International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 11, Volume 2.

[16] K Arutchelvan Dr R Periyasam. 2015. Cancer Prediction System Using Data Mining Techniques, International Research Journal of Engineering and Technology (IRJET) Volume: 02 Issue: e-ISSN:2395-00569.

[17] Muhammad Sufyian Bin Mohd Azmi ZaihismaChe Cob. 2010. Breast Cancer Prediction Based on Backpropagation Algorithm IEEE Student Conference on Research and Development.

[18] Nimna Jeewandara , PPG Dinesh Asanka. November 2017. Data Mining Techniques In Prevention And Diagnosis Of Non Communicable Diseases, International Journal Of Research In Computer Applications And Robotics ISSN 2320-7345, Vol.5 Issue 11, Pg.: 11-17.

[19] P.Ramachandran N. Girija T. Bhuvaneswar. 2014. Early Detection and Prevention of Cancer Using Data Mining Techniques, Internaiton Journal of Computer Applications (0975 - 8887) Volume 97-No.13.

[20] Pallavi Mirajkar, G. Prasanna Lakshmi. 2017. Analysis and Identification of Cancerous Gactors Using Rule Based Classifier of Data Mining Techniques International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835 Volume-4, Issue-6.

[21] Pallavi Mirajkar, Dr. G. Prasanna Lakshmi. 2018.An Integrated Cancer Prediction System Using Data Mining Techniques, International Journal of Scientific Research in Computer Science, Engineering and Information Technology IJSRCSEIT , Volume 3 ,Issue 1 ,ISSN : 2456-3307.

[22] Parshva Jain, Rahul Vijayvargiya & Arsheen. November, 2017. Use Of Machine Learning Techniques To Effectively Manage And Diagnose Huge Amount Of Data In The Field Of Health Care Industry, IJESRT International Journal Of Engineering Sciences & Research Technology.

[23] R. Rajbharath , L. Sankari. April 2017. Predicting Breast Cancer using Random Forest and Logistic Regression, IJESC, International Journal of Engineering Science and Computing.

[24] Rozita Jamili Oskouei, Nasroallah Moradi Kor, and Saeid Abbasi Maleki. March 1, 2017.Data mining and medical world: breast cancers' diagnosis, treatment, prognosis and challenges, American Journal of Cancer Research, ISSN: 2156-6976; 7(3): 610–627.

[25] Seyed Mohammad Jafar Jalali, Sérgio Moro, Mohammad Reza Mahmoudi, Keramat Allah Ghaffary, Mohsen Maleki, Aref Alidoostan,Repositório.2017. A Comparative Analysis of Classifiers in Cancer Prediction Using Multiple Data Mining Techniques, ISCTE-IUL.

[26] Shelly Gupta Dharminder Kumar, Anand Sharma. 2011. Data Mining Classification Techniques Applied Forbreast Cancer Diagnosis And Prognosis, Indian Journal of Computer Science and Engineering (IJCSE) ISSN : 0976-5166 Vol. 2.

[27] Shweta Kharya.2012. Using data mining Techniques for diagnosis and prognosis of cancer disease,Internatioanl Journal of computer Science,Engineering and Information Technology(IJCEIT),Vol.2.

[28] Shakuntala Jatav and Vivek Sharma. February 2018. An Algorithm For Predictive Data Mining Approach In Medical Diagnosis, International Journal of Computer Science & Information Technology (IJCSIT) Vol 10, No 1.

[29] V.Krishnaiah, Dr G Narsimha Dr N Subhash Chandre. 2013. Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques, International Journal of Computer Science and Information Technologies, Vol. 4 (1), 39 – 45.

[30] Yao Liu and Yuk Ying Chung.2011. Mining Cancer data with Discrete Particle Swarm Optimization and Rule Pruning 978-1-61284-704-7/11/IEEE.