ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 6, November 2016

A Survey on Future Trends of Data Mining in Predicting Various Cancer in Medical Healthcare System

S Swarajyam, M Ramesh, S Meena

Abstract*— One of the fastest growing fields is health care industries. The healthcare industry collects immense amounts of medical data. These data are collected from the patients who have undergone any kind of medical treatment or tests. By mining into these data, hidden patterns and relationships can be discovered for efficient analysis, diagnosis and prognosis. If this information gathered is aptly utilized then a system can be generated to assist the medical practitioner to take medical decisions. The apparent relationship that has been discovered with respect to cancer does not give accurate results when applied to prediction models. Thus we need to discover new relationships and patterns which will help set up a more accurate decision support system. Data mining is a process of getting hidden patterns from the dataset. Various data mining techniques are clustering, classification, association analysis, regression, summarization, time series analysis and sequence analysis, etc. The objective of this paper is to review the past work done on the prediction of lung, breast and liver cancer, three of the most fatal diseases. The aim of this work is to provide a succinct and concise overview of the work done in this field.

***Index Terms*— Breast Cancer, Lung Cancer, Liver Cancer, Data Mining, Classification Rules.**

I. INTRODUCTION

Cancer is the name given to a collection of related diseases. In all types of cancer, some of the body's cells begin to divide without stopping and spread into surrounding tissues. Cancer can start almost anywhere in the human body, which is made up of trillions of cells. Normally, human cells grow and divide to form new cells as the body needs them. When cells grow old or become damaged, they die, and new cells take their place. When cancer develops, however, this orderly process breaks down. As cells become more and more abnormal, old or damaged cells survive when they should die, and new cells form when they are not needed. These extra cells can divide without stopping and may form growths called tumors. Lung cancer is the second most common cancer, accounting for about one out of five malignancies in men and one out of nine in women and the leading cause of cancer death among both men and women where about 1 out of 4 cancer deaths are from lung cancer. Each year, more people die of lung cancer than of colon, breast, and prostate cancers combined. Liver cancer is the sixth most common cancer in the world. Globally, hepatocellular carcinoma (HCC) is among the most prevalent malignant tumors. Worldwide, over a million deaths per year (about 10% of all deaths in the adult age range) can be attributed to hepatocellular carcinoma. Liver cancer is usually a life-threatening condition. However, like lung cancer, it may be effectively treated if found early. Breast cancer is cancer that develops from breast tissue. Signs of breast cancer include a lump in the breast, a change in breast shape, dimpling of the skin, fluid coming from the nipple, or a red scaly patch of skin. In those with large spread of the disease, there may be any of the following: bone pain, swollen lymph nodes, shortness of breath, or yellow skin. Over the years, many researchers have been trying to create a model which can help efficiently diagnose the possibility of cancer, as the earlier it is diagnosed the better are the chances of survival. This paper aims to explore the many recent works published on the matter and provide a coherent and collected summary of such work that is based on the prediction of possibility of cancer based on features of the patient which can be known without any invasive medical procedures. The various papers that we have reviewed here basically make use of either of the following algorithms in various combinations: genetic algorithm, artificial neural network and fuzzy c means. The father of the original Genetic Algorithm was John Holland who invented it in the early 1970's. Genetic Algorithms are adaptive heuristic search algorithms based on the evolutionary ideas of natural selection and genetics. Genetic algorithms harness the power of evolution to solve optimization problems. So different processes of natural selection like recombination and mutation are incorporated into the algorithm. The father of the original Genetic Algorithm was John Holland who invented it in the early 1970's. Artificial neural networks are computer programs designed to simulate the way in which the human brain processes information.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 6, November 2016

ANNs gather their knowledge by detecting the patterns and relationships in data and learn over time through experience and not from programming. It basically refers to a large network of processing elements, which behave like neurons, arranged in multiple layers. Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to be a part of two or more clusters. This method (developed by Dunn in 1973and improved by Bezdek in 1981) is frequently used in pattern recognition. This sort of clustering is most apt for fuzzy datasets. We believe these three methods are few of the most efficient techniques to be applied in the prediction of the possibility of cancer.

II. LITERATURE REVIEW

K. Balachandaran et al. [1] have an interesting approach towards prediction of lung cancer. In this paper, the given dataset‘s dimensionality is reduced using Artificial Bee Colony (ABC) algorithm and the reduced dataset containing just the high risk factors and symptoms which cause lung cancer are fed into the Feed Forward Back Propagation Neural Network (FFBNN). While training, the FFBNN parameters are optimized using ABC algorithm. During the testing process, more number of patient‘s data is given to well-trained FFBNN-ABC to validate whether the given testing data predict the lung disease perfectly or not. The accuracy of the proposed technique is 90% and the sensitivity of the same is 88% while the specificity is 100%.

K. Polat et al. [2] have detected lung cancer using principles component analysis (PCA), fuzzy weighting preprocessing and artificial immune recognition system (AIRS). The system has three stages. First, dimensionality of lung cancer dataset that has 57 features was reduced to four features using principles component analysis. Second, a weighting scheme based on fuzzy weighting pre- processing was utilized as a pre-processing step before the main classifier. Third, artificial immune recognition system was used classifier. Experiments were conducted on the lung cancer dataset to diagnose lung cancer in a fully automatic manner. The obtained classification accuracy of system was 100% and it was very promising with regard to the other classification applications.

V. Krishnaiah et al.[3] discusses the statistically significant effect of symptoms and risk factors in pre-diagnosis stage. A prototype lung cancer disease prediction system has been developed using data mining classification techniques which extracts hidden knowledge from a historical lung cancer disease database. The healthcare industry amasses large amounts of medical data which are rarely properly exploited to discover hidden patterns and relationships. For data preprocessing and effective decision making. One Dependency Augmented Naïve Bayes classifier, also known as ODANB, and Naive Creedal Classifier 2, also known as NCC2, are used. This appears to be an extension of Naïve Bayes to imprecise probabilities that aims at delivering robust classifications also when dealing with small or incomplete data sets. According to the author’s experimental results, the most effective model to predict patients with lung cancer disease appears to be Naïve Bayes followed by IF-THEN rule, Decision Trees and Neural Network.

Parag Deoskar, et al. [4] proposes to combine data mining and ant colony optimization techniques for appropriate rule generation and classification, which can lead to accurate cancer classification. In addition to this, it provides basic framework for further improvement in medical diagnosis. This paper is divided into sections which handle the following topics each: medical data mining; ant colony optimization; related works; theoretical extraction. It is seen that ant colony optimization helps in increasing the prediction (of the disease) value significantly. The authors provided future suggestions like application of neural network and Fuzzy based technique to train cancer data set for finding better classification, applying optimization techniques like ACO to improving the detection, use of machine learning environment or Support Vector machine and the use of homogeneity based algorithm to find over fitting and over generalization Characteristics.

P. Ramachandaran et al. [5] uses data mining technology such as classification, clustering and prediction to identify potential cancer patients. The gathered data is preprocessed to yield significant patterns using decision tree algorithm which is then clustered using K- means clustering algorithm to separate cancer and non-cancer patient data. The cancer cluster is further subdivided into six clusters. Finally a prediction system is developed to analyze risk levels which help in prognosis. This research helps in detection of a person‘s predisposition for cancer before going for clinical and lab tests which is cost and time consuming. The model shows an accuracy of 99.87%.

Thangaraju P et al. [6] proposed a system is to find out the medical issues of Lung cancer and find out the stages of

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 6, November 2016

the lung cancer patients by using the data of Patients Details and risk factors of lung cancer which are collected from the hospital database. Mainly decision tree is used for predicting the Lung Cancer Disease from the given dataset instances. In the proposed method mainly decision tree is used for predicting the Lung Cancer Disease from the given data set instances and the proposed model contains three different types of decision tree algorithms such as Naive Bayes, Decision Table and j48 are applied on type Lung Cancer Disease dataset in the WEKA tool and the performance is calculated. In this paper, the Naive Bayes classified 253 instances and produce the 83.4% of accuracy for prediction of lung cancer while the Decision table classified 231 instances and produced 76.2% of accuracy and the J48 classified 235 instances and produced 77.5% of accuracy.

Kawsar Ahmed1, et al. [7] proposed to significant pattern prediction tools for a lung cancer prediction system was developed. The lung cancer risk prediction system should prove helpful in detection of a person‘s predisposition for lung cancer. The early prediction of lung cancer should play a pivotal role in the diagnosis process and for an effective preventive strategy. In the initial stages, 400 cancer and non-cancer patients data were collected from different diagnostic centers, pre-processed and clustered using a K-means clustering algorithm for identifying relevant and non-relevant data. Next the significant frequent patterns are discovered using Apriority and a decision tree algorithm. Next the experimental results are separated into two sections where one is the discovery of significant frequent patterns and another is the representation of prediction tools for Lung Cancer. Using the data from data warehouse, the significant patterns are extracted for Lung cancer prediction. The collected data are pre-processed by deleting the duplicate records and adding the missing values. Then pre-processed data is clustered using K-means cluster algorithm with k equal to 2.

T. Sowmiya et al [8] speaks of the urgent need for early detection of the cancer that can save the life and help the survivability of the patients who affected by this diseases. This paper surveys several aspects of data mining procedures which can be used for lung cancer prediction of the patients. It reiterates the importance of data mining concepts in lung cancer classification. It also reviews the aspects of ant colony optimization (ACO) technique in data mining. The paper examines the compromises in selection and dimensionality reduction and showed that acceptable plans could be obtained in approximately 30 minutes. ROCO strategies satisfy all of the clinical restrictions that were satisfied by the planner‘s plans; with the same PTV D95, there were no significant differences between the OAR sparing achieved by ROCO and the organ sparing achieved by the medical plans. The paper assures that ROCO will be flexible enough for general external beam radiation remedy preparation, and is not confined to simpler treatments such as prostate cancer. A major improvement made to ROCO in the current work is the incorporation of ROCO into MSKCC‘s clinical treatment scheduling system. ROCO none seems to be capable of evaluation and inscription beam and dose information directly to/from the treatment scheduling system. This case study assorted data mining and ant colony optimization techniques for appropriate rule generation and classifications on diseases, which pilot to exact Lung cancer classifications. In additionally to, it provides basic framework for further improvement in medical diagnosis on lung cancer.

Using Data Mining Techniques

Miss Jahnavi Joshi et al. [9] developed a new sample model for diagnosis breast cancer patients. There are thirty seven classification rules are used. By comparing the rules the model has been developed. Using this model, it is seen that the patterns of the dataset can be made efficiently. Dataset is figured by using WEKA mining tool. After that, those classification algorithms are used on that dataset. Then the sample evaluation is done on the healthy and sick patients and the results are given to the predictive classifier to discover the pattern. Web mining can be classified into three categories which are structure, usage, and control. Above mentioned three categories the usage is used for this model, classification rules are applied on these dataset. Some of them of those classifier rules are Bayes Net, SGD, Decision table, Decision Stump, SMO, Multi-Scheme, LMT, Voted, Random-Committee, Random-Forest, IBK etc., then prototype is evaluated in order to determine healthy and sick people. By this approach it is seen that discovery of patterns can be efficient. This method is useful to discover the hidden patterns and helps the doctors and medical practioners to take the medical decision. The proposed model can identify the the type of the breast cancer. This model can make the generic model for different areas like commercial model, electricity model.

Ibrahim M. El-Hasnony et al. [10] presented a system to classify the breast cancer. This system is combined of three methods. In order to pre-process the data FRFS (fuzzy rough feature selection) is used to handle the data

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 6, November 2016

which are missed. To make the data cluster clustering algorithms used and features are reduced by the fuzzy rough feature selection and also the features which are reduced is merged. The classification of data is done by the D-KNN (discernibility nearest neighbor) classifier. At last the performance is evaluated. The data set is taken from the UCI repository and this model is examined under that dataset. By using K-means clustering algorithm with k-value 2 the dataset are pre-processed for noise containing data and missing data, for this process WEKA tool and miner (rapid) is used for clustering utilization. The reduction of the clustered data is done by that selection algorithm. The reduced features are combined together to form the new dataset.Atlast,the classification is achieved by the classifier(D-KNN).This model can classify the instances of the new dataset with accuracy up to 98%.The classification accuracy can be increased efficiently by this model.

RonakSumbaly et al. [11] developed the model by using data mining methods for the diagnosis of breast cancer patients to apply the treatment and three other methods like mammography(digital),Naive Bayes model, Neural networks are presented to make the comparison with the proposed model.Bythis,the decision tree model is constructed. The dataset applied in this proposed model is Wisconsin Breast cancer datasets which are taken from UCI. The preprocessing of data is done by J48 decision tree data mining method and after that the data is given to WEKA data mining tool for analysis-fold cross validation (where the value of k is 10) method is applied to form the training data. The tree is constructed and the leaf nodes of the tree determine whether the cancer is malignant or benign. The tree is represented level-wise when WEKA mining tool is applied on that preprocessed dataset. Fourteen leaves (leaf nodes) are generated by that tree and the number of total tree was twenty four. This model is tested over 699cases and it gives high accuracy and significant result in most of the cases. The accuracy of prediction of this model is 98%.

LIVER CANCER

Fabio Bagarell et al[12] examined if an Artificial Neural Network is capable in detection of hepatobiliary disease amongst certain patients with knownhepatobiliary diseases, using only medical and few laboratory findings, to construct a tool for early and ―pre-imaging‖ diagnosis of patients. Medical records of 270 patients were considered. ANN can extract most similar case from database in order to deal with new problems. Each neuron has multiple input layer but only one output layer. Software used is Easy NN-Plus. The end result showed an accuracy of 96%. This method reduced diagnostic errors and built a cost efficient way of handling medical resources.

Herng-Chia Chiu et al[13] constructed prediction models based on medical records for disease free survival using a database for hepatocellular carcinoma (HCC) patients who had received hepatic resection. Survival was defined as disease-free survival after 1, 3, or 5 years. The presence of an event (death or recurrence) was coded as 1, and absence of an event (disease-free survival) was coded as 0. The input layer in each of the three models comprised of 17 neurons: age, gender, liver cirrhosis, chronic hepatitis, AST, ALT, total bilirubin, albumin, creatinine, ASA classification, Child-Pugh classification, TNM stage, tumor number, portal vein invasion, biliary invasion, surgical procedure, and post-operative complication. In the hidden layers, the numbers of neurons were optimized by training and validating data in a trial-and-error process to maximize predictive accuracy. Only one neuron was obtained as an output in all the three cases representing the disease-free survival. The ANN model overpowered the LR and DT models in terms of prediction accuracy.

Md. Osman GoniNayeem et al [14] suggested that ANN turns out to be the most vital classification approach on considering three different diseases (heart disease, liver disorder, lung cancer). Feed-forward back propagation neural network algorithm with Multi-Layer Perceptron (MLP) is used as a classifier to distinguish between infected or non-infected person. MLP is a feed forward artificial

Neural network model used to maps input data onto appropriate outputs. A MLP consists of multiple layers of nodes in a directed graph; each layer is entirely connected to the next one. The results of applying the ANNs methodology to diagnosis of these diseases based upon selected symptoms show abilities of the network to learn the patterns corresponding to symptoms of the person. Here in case of liver disorder prediction patients are classified into four categories: normal condition, abnormal condition (initial), abnormal condition and severe condition. ANN has the ability to learn complex and nonlinear relationships including noisy or less precise information. For liver disorder and lung cancer prediction networks shows an accuracy of 82% and 91% respectively.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 6, November 2016

Zhang et al [15] explored the factors affecting liver cancer recurrence after hepatectomy.The BP algorithm was used to perform the prognosis on the selected statistical information. Eighteen factors were selected by unifactor analysis out of which nine factors were selected by multi-factor analysis. The nine factors selected can be as important indexes to evaluate the recurrence of liver cancer. The ANN is a better approach to evaluate clinical data. The study can provide the basis with scientific and objective data for analyzing prognosis of liver cancer. The statistic method used in this paper is maximum likelihood estimate. This research was supported by NSFC.

Joseph A. Cruz et al [16] intended to identify the types of machine learning methods being used, the types of training data being constructed, and the kinds of cancers being studied and the overall performance of these methods in predicting cancer susceptibility. Although in the recent studies it has been noted that ANN has outperformed most of the machine learning languages yet there are still other alternative strategies to be developed. When dealing with cancer three primary factors need to be examined namely it‘s prediction, recurrence and survivability. It is clear that machine learning methods tend to improve the performance or predictive accuracy of most prognoses, especially when compared to conventional statistical or expert-based systems. The only limitation being that the whole study is based on assumptions and cross examination so the initial validation has to be done with utter care and has to be crucially examined. You can add the remaining content as it is but the heading must be Time New Roman Front of size 11 with bold and the content must be as of introduction i.e time new roman of size 10 and must be justified alignment.

III. CONCLUSION

In conclusion, the current compilation of several paper works can be described as a preliminary study. It will need further validation in a separate cohort of patients with lung, breast and liver problems. In fact, it is clear that a similar ANN can be organized for different kind of diseases; so many possibilities were opened by our analysis. Our proposed method is using a hybrid Artificial Neural Network and Genetic Algorithm for classification which works on the clusters of Fuzzy C Means. In this review, the focus is on the current research being carried out using the data mining techniques to enhance the disease(s) forecasting process.

ACKNOWLEDGMENT

There is no scope for learning and improvement unless one makes mistakes. We take this opportunity to express our profound gratitude to everyone who has extended a helping hand to us in this endeavor, no matter what their contribution has been. We shall keep working on this topic to further the cause of cancer prediction in the early stages so as to save lives that need not be lost unnecessarily.

REFERENCES

[1] Balachandran, DR. R. Anitha, ―An efficient optimization based lung cancer prediagnosis system with aid of feed forward back propagation neural network(FFBNN)‖, Journal of Theoretical and Applied Information Technology.

[2] Kemal Polat and Salih Gunes, ―Principles component analysis, fuzzy weighting pre- processing and artificial immune recognition system based diagnostic system for diagnosis of lung cancer‖, Expert Systems with Applications.

[3] Krishnaiah, Dr.G.Narsimha, Dr.N.Subhash Chandra, ―Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques‖, (IJCSIT) International Journal of Computer Science and Information Technologies.

[4] Parag Deoskar, Dr. Divakar Singh, Dr. Anju Singh, ―Mining Lung Cancer Data and Other Diseases data using Data Mining Techniques.

[5] Ramachandran, N.Girija, T.Bhuvaneswari, ―Early Detection and Prevention of Cancer using Data Mining Techniques‖, International Journal of Computer Applications (0975 – 8887) Volume 97– No.13, July 2014.

[6] Thangaraju P, Barkavi G, Karthikeyan T, ―Mining Lung Cancer Data for Smokers and Non- Smokers by Using Data Mining Techniques‖, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 7, July 2014.

[7] Kawsar Ahmed, Abdullah-Al-Emran, Tasnuba Jesmin, Roushney Fatima Mukti, Md Zamilur Rahman, Farzana Ahmed, ―Early Detection of Lung Cancer Risk Using Data Mining‖, Asian Pacific Journal of Cancer Prevention.

[8] T. Sowmiya, M. Gopi, M. New Begin Thomas Robinson, ―Optimization of Lung Cancer using Modern data mining techniques.‖, International Journal of Engineering Research.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 5, Issue 6, November 2016

[9] Miss Jahanvi Joshi,Mr. Rinal Doshi,Dr. Jigar Patel, "Diagnosis and Progonosis breast cancer using classification rules", International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014.

[10] Ibrahim M. El-Hasnony, Hazem M. El-Bakry, Ahmed A. Saleh,"Classification of Breast Cancer Using Soft computing Techniques", International Journal of Electronics and Information Engineering, Vol.4, No.1, Mar 2016.

[11] Ronak Sumbaly N. Vishnusri. S. Jeyalatha ―Diagnosis of Breast Cancer using Decision Tree Data Mining Technique", , International Journal of Computer Applications (0975 – 8887) Volume 98– No.10, July 2014.

[12] Prof.Fabio Bagarello, Pasque Lemansueto,"Artificial Neural Networks in Liver Cancer: An economic and pre-imaging Diagnosis". Maecellocam arata, Italy. Published in 2013.

[13] Wen-Hsien Ho, King-Teh Lee, Hong-Yaw Chen, Te-Wei Ho, Herng-Chia Chiu,"Artificial Neural Network to explore effecting factors of Hepatic Cancer recurrence". Published January 3, 2012

[14] Md. Osman Goni Nayeem, Maung Ning Wan, Md. Kamrul Hasan, ―Prediction of Disease Level Using Multilayer Perceptron of Artificial Neural Network for Patient Monitoring ,International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-5 Issue-4, September 2015

[15] Xian-min,Zhang,Zhi-jian,," The method of artificial neural network applied to explore the effecting factors of hepatic cancer recurrence after hepatectomy ",China.

[16] Joseph A. Cruz, David S. Wishart, ―Applications of Machine Learning in Cancer Prediction and Prognosis‖,Departments of Biological Science and Computing Science, University of Alberta Edmonton, AB, Canada.

AUTHOR BIOGRAPHY

First Author: Name – Siddamalla Swarajyam, Assistant Profess, MLRIT, Hyderabad.

Second Author: Name – Munipala Ramesh, IT Analyst, TCS, Hyderabad.

Third Author: Name – Siddamalla Meena, Student, GNIT, Hyderabad.