Asian Journal of Computer Science and Technology
ISSN: 2249-0701 Vol.8 No.S2, 2019, pp. 52-56

A Study: Breast Cancer Prediction Using Data Mining Techniques

B. Gousbi[1] and A. R. Mohamed Shanavas[2]
[1]M.Phil Research Scholar, Department of Computer Science,
[2]Associate Professor, Department of Computer Science& Information Technology,
[1&2]Jamal Mohamed College, Tiruchirappalli, Tamil Nadu, India
E-Mail: gousbibazir@gmail.com

***Abstract* - Data mining is the extraction of unseen predictive info from huge databases, is the process of arranging through enormous data sets to recognize patterns and create relationships to resolve the problems through data analysis. Cancer is one of the primary reasons of death wide-reaching. Timely detection and prevention of cancer plays a very vital role in decreasing deaths affected by cancer. Identification of genetic and environmental factors is very significant in emerging novel methods to identify and avert cancer. Many researchers' use data mining techniques like clustering, classification and prediction find potential cancer patients. This paper focuses on a breast cancer prediction system built on data mining techniques. With the help of this system, people can guess the possibility of the breast cancer in the former stage itself.**

Keywords: **Data Mining, Breast Cancer, Prediction, Classification and Clustering**

I. INTRODUCTION

Data mining is the method of determining possibly interesting, useful, and formerly unknown patterns of a huge group of data. Data mining is a multidisciplinary field, drawing work from areas including artificial intelligence, machine learning, statistics, pattern recognition, information retrieval, database technology, neural networks, high-performance computing, knowledge-based systems, and data visualization. We present techniques for the encounter of patterns concealed in large data sets, directing on issues relating to their effectiveness, feasibility, scalability and usefulness. Data mining is a process to extract the implied information and knowledge which is possibly useful and people do not know in advance, and this extraction is from the fuzzy, mass, noisy, incomplete, and random data.

Data mining is the use of automatic data analysis techniques to discover previously hidden relationships between data items. Data mining often includes the analysis of data deposited in a data warehouse. Three of the major data mining techniques are clustering, regression and classification. Data Mining, also usually well-known as Knowledge Discovery in Databases (KDD), denotes to the nontrivial withdrawal of unsaid, formerly unidentified and probably useful information from data in databases [1]. Though data mining and knowledge discovery in databases (or KDD) are often treated as substitutes, data mining is essentially part of the knowledge discovery process.

In the prediction of cancer, the Data Mining techniques are applied together to build a novel method to analyze the presence of cancer for a particular patient. While beginning to work on a data mining problem, it is first essential to bring all the data collected into a set of instances [2]. Assimilating data from dissimilar sources commonly present numerous challenges. The data must be collected, combined, and cleaned up. Only then it can be used for handling over machine learning techniques.

Cancer is one of the utmost communal diseases in the world that outcomes in common of death. Cancer is initiated by the unrestrained development of cells in any of the tissues or parts of the body. Cancer may arise in any part of the body and may extend to numerous other parts. Only timely finding of cancer at the beginning stage and avoidance from distribution to other parts in malignant stage could save a person's life.

Cancer is a possibly deadly disease produced mostly by environmental factors that change genes encoding critical cell-regulatory proteins. The subsequent aberrant cell behavior leads to extensive masses of abnormal cells that abolish nearby normal tissue and can spread to energetic organs, resulting in scattered disease [3], usually a harbinger of imminent patient death. More pointedly, globalization of unnatural lifestyles, mainly cigarette smoking and the acceptance of numerous features of the modern Western diet (high fat, low fiber content) will raise cancer occurrence.

II. RELATED WORKS

V. Krishnaiah *et al.,* [4] established a prototype lung cancer disease prediction system using data mining classification techniques. The best effective model to guess patients with Lung cancer disease seems to be Naïve Bayes followed by IF-THEN rule, Decision Trees and Neural Network. For Analysis of Lung Cancer Disease Naïve Bayes perceives enhanced results and managed better than Decision Trees.

Sahar A. Mokhtar *et al.,* [5] have studied three different classification models for the guess of the harshness of breast masses specifically the artificial neural network, decision tree and support vector machine. The decision tree model is built using the Chi-squared automatic interaction detection method and pruning method was used to discover the

optimal structure of artificial neural network model and lastly, support vector machine have been constructed using the polynomial kernel. The acts of the three models have been assessed using statistical measures, gain and Roc charts. Support vector machine model beats the other two models on the expectation of the severity of breast masses.

Ada *et al.,* [6] made an effort to identify the lung tumors from the cancer images and supportive tool is established to check the usual and unusual lungs and to guess survival rate and years of an abnormal patient so that cancer patients' survives can be protected.

Charles Edeki *et al.,* [7] recommends that none of the data mining and statistical learning algorithms applied to breast cancer dataset beat the others in such way that it could be stated the optimal algorithm and none of the algorithm accomplished poorly as to be removed from a future expectation model in breast cancer survivability tasks.

Zakaria Sulimanzubi *et al.,* [8] used some data mining techniques such as neural networks for discovery and classification of lung cancers in X-ray chest films to categorize difficulties aiming at finding the features that specify the group to which every case belongs.

III. CLUSTERING TECHNIQUES

Clustering[9]is the task of splitting the population or information focuses into several gatherings with the end aim that information emphases in related gatherings are more like other information focuses in a related gathering than those in dissimilar gatherings. In simple words, the point is to separate clusters with proportional attributes and dole out them into collections. A typical fascinating assignment in which one tries to identify a narrow organization of classifications or group to portray the information is known as clustering. There are different types of clustering methods or techniques available in data mining. They are listed here:

A. Hierarchical Methods:It follows two methods which are bottom-up and top-down where all clusters are united and are converted into one andall interpretations are divided into dissimilar bunches respectively.

TABLE I ADVANTAGES AND DISADVANTAGES OF HIERARCHICAL METHODS

Advantages	1. Embedded flexibility about the level of granularity 2. Ease of handling of any forms of resemblance or distance 3. Consequently, applicability to several attribute types
Disadvantages	1. Imprecision of termination criteria 2. The point that most hierarchical algorithms do not reconsider once constructed (intermediate) clusters with the purpose of their development

B. Grid-Based Methods: The objects organized shape a framework. The object form is quantized into an inadequate number of cells that shape a network arrangement.

TABLE II GRID BASED ALGORITHMS AND ITS ADVANTAGES

Algorithms	Advantages
Wave Cluster	1. Great quality of clusters 2. Facility to work well in comparatively high dimensional spatial data 3. Successful handling of outliers
ASGC	1. Short time complexity 2. It is a non-parametric algorithm 3. It preprocesses the data space and decreases the dimension of the data space

C. Partitioning Methods: Assume we are given a database of "n" items and the partitioning strategy constructs "k" segment of information. Every division will have a cluster and k <= n. It indicates that it will organize the information into k gatherings, which satisfy the supplementary prerequisites.

1. Every gathering comprises no less than one object.
2. Every question must have a place with exactly one gathering.

TABLE III PARTITIONING ALGORITHMS AND ITS ADVANTAGES

Algorithm	Advantages
K-Mean	1. Quicker than hierarchical clustering methods 2. Creates tighter clusters than Hierarchical Clustering Method
K-Medoid	1. Simple to realize and easy to implement 2. Fast and converges in a static number of steps 3. Fewer sensitive to outliers than other partitioning algorithms
CLARA	1. CLARA Algorithm deals with higher data sets than PAM (Partition Around Medoids)
CLARANS	1. Stress-free to handle outliers 2. CLARANS result is more the active as relate PAM and CLARA

D. Density Based Methods:This method is centered on two functions types that are connectivity and density functions. The simple idea behind it is that the thick clusters are made and they should develop as long as they cross the threshold of the nearest clusters.

TABLE IV DENSITY BASED ALGORITHMS AND ITS ADVANTAGES.

Algorithms	Advantages
DBSCAN	1. Catch out clusters of unfamiliar shapes. 2. Handle noise and outliers 3. One Scan method.
OPTICS	1. It does not wantdensity parameters. 2. Clustering order is useful in abstract the simple clustering information.
DENCLUE	1. Quicker than existing algorithms. 2. Good for database that Comprisesvast amount of noise. 3. Create precise result.

Initially, all data objects inside the group are equally density related to each other. Secondly, if a data object is density related to any other data object inside the group then both the data objects must be the portion of the similar group.

E. Model Based Methods: In this model is theorized for every group to localize the greatest shape of data for a given model.

IV. CLASSIFICATION TECHNIQUES

Constructing accurate and effective classifiers for huge databases is one of the vital tasks of data mining and machine learning research. Constructing effective classification systems is one of the significant tasks of data mining. Several different types of classification techniques include Decision Trees, Naïve Bayesian methods, Neural Networks, Logistic Regression , SVM and KNN etc. [11].

A. Decision Tree: Decision tree models are usually used in data mining to observe the data and make the tree and its rules that will be used to make predictions. The decision tree is a classifier in the form of a tree structure where every node is also a leaf node, specifying the value of the aim attributes or class of the examples, or a decision node, stating some test to be carried out on a single attribute-value, with one branch and sub-tree for every possible result of the test.

B.Neural Networks: Neural networks are talented of forming very complex, typically non-linear functions. It is made up of a structure or a network of several interrelated units (artificial neurons). Each of these units contains of input/output characteristics that implement a confined computation or function.

C. Naive Bayes: The Naive Bayes is a rapid method for the establishment of statistical predictive models. NB is based on the Bayesian theorem. This classification technique examines the association between every attribute and the class for every instance to develop a conditional probability for the associations between the attribute values and the class.

D. Logistic Regression: LR is considered as the standard statistical method for forming binary data. It is an enhanced alternate for a linear regression which gives a linear model to each of the class and guesses hidden instances basing on the popular vote of the models.

E. Support Vector Machine: SVMs are a set of associated supervised learning methods that examine data and identify patterns, used for classification and regression analysis. SVM is an algorithm that tries to discover a linear separator (hyper-plane) among the data points of two classes in multidimensional space. SVM denotes a learning technique which follows the principles of statistical learning theory.

F. K-Nearest Neighbor: K-Nearest Neighbor (KNN) classification categorizes instances based on their comparison. An object is categorized by a majority of its neighbors. K is constantly a positive integer. The neighbors are nominated from a set of objects for which the exact classification is recognized. The training examples are defined by n dimensional numeric attributes. Every sample signifies a point in an n dimensional space. In this way, all of the training samples are deposited in a dimensional pattern space.

V. COMPARATIVE STUDY OF VARIOUS PROPOSALS

A. Novel Multi Layered Method

P. Ramachandran *et al.,* [2] proposed a new model work in which the composed data are pre-processed and deposited in the knowledge base to construct the model. Seventy five per cent of the total data are taken as the training set to construct the classification and clustering model the residual of which is taken for challenging purpose. The decision tree model is constructed using the classification rules, the significant frequent pattern and its equivalent weights. The clustering model is constructed by using the k-means clustering algorithm. The model is then tested for accurateness, sensitivity and specificity using test data laterally with integration it to the knowledge base. Lastly the model is calculated using Support Vector Machine.

B. Cancer Prediction System

K. Arutchelvan *et al.,* [3] proposed an architecture, in which data mining technique centered cancer prediction system relating the prediction system with mining technology was used. In this model, the authors have used one of the classification algorithms called decision tree. When the user comes into the cancer prediction system, they want to reply the queries, allied to genetic and non-genetic factors. At that time the prediction system gives the risk value to every question based on the user replies. When the risk value is anticipated, the kind of the risk can be determined by the prediction system. The prediction system has four levels of risk like low level, intermediate level, high level and very high level. Based on the expected risk values the kind of risk will be allocated.

C. Adaptive Neuro Fuzzy Inference System (ANFIS)

C. Kalaiselvi *et al.,* [10] proposed a new Adaptive Neuro Fuzzy Inference System (ANFIS). The multi factorial, long-lasting, severe diseases like diabetes and cancer have difficult connection. Once the glucose level of the body goes to unusual level, it will lead to Kidney failure, Blindness, Heart disease and also a Cancer. Epidemiological studies have verified that numerous cancer varieties are conceivable in patients having diabetes. Several researchers have proposed methods to detect diabetes and cancer. Adaptive Neuro Fuzzy Inference System (ANFIS) which is

developed to increase the classification exactness and to attain improved efficiency.

D. Support Vector Machine

G. Ravi Kumar *et al.,* [11] targets to createan exact classification model for Breast cancer prediction, in order to make complete use of the priceless information in clinical data, particularly which is commonly ignored by most of the surviving methods when they aim for great prediction accuracies. They have completed experiments on WBC data. The dataset is separated into a training set with 499 and test set with 200 patients. In this research, they relate six classification techniques in Weka software and evaluation results indicate that Support Vector Machine (SVM) has greater prediction accuracy than those methods. Different methods for breast cancer uncovering are discovered and their accuracies are related. With these results, they conclude that the SVM is more appropriate in handling the classification problem of breast cancer prediction, and they endorse the use of these methods in parallel classification problems.

E. Simple Logic Classification

Vikas Chaurasia *et al.,* [12] presents an analysis system for detecting breast cancer centered on RepTree, RBF Network and Simple Logistic. On trial stage, 10-fold cross validation process was smeared to the University Medical Centre; database to assess the proposed system enactments. The greatest algorithm based on the patient's data is Simple logistic Classification with accurateness of 74.47%. The exact classification rate of the proposed system is 74.5%. This research established that the Simple Logistic can be used for decreasing the aspect of feature space and proposed Rep Tree and RBF Network model can be used to get fast spontaneous diagnostic systems for other diseases. The result of this paper suggests that amongst the machine learning algorithm tested, Simple logistic classifier has the latent to knowingly increase the conventional classification methods used in.

F. Hybrid Classification Algorithm

K. Sivakami *et al.,* [13] suggested a hybrid classification algorithm for breast cancer patients, which incorporate DT and SVM algorithms. The proposed algorithm was self-possessed of two main phases. The first phase is Information treatment and option abstraction followed by DT-SVM hybrid model predictions. The Classification had two chief phases, Training and Testing phases. The input parameters for SVM were enhanced using DT algorithm.

The SVM algorithm was used to categorize breast cancer patients into one of two classes (Benign/Malignant).In assessment of data mining techniques, this research used precision indicator to assess classification effectiveness of different algorithms. The proposed algorithm was associated with different classifier algorithms using Weka tool.

TABLE V COMPARATIVE STUDY OF VARIOUS PROPOSED WORKS

Author	Method	Algorithm	Accuracy	Evaluation
P.Ramachandran *et al.,*	Novel Multi Layered Method	Decision tree k-means clustering algorithm	99.866%	Weka
K. Arutchelvan *et al.,*	Cancer Prediction	Decision tree	Better than existing	Weka
C. Kalaiselvi *et al.,*	Adaptive Neuro Fuzzy Inference System (ANFIS)	k-means Algorithm	66-77%	Weka
G. Ravi Kumar *et al.,*	Classifier algorithm	Support Vector Machine(SVM)	97.59%	Weka
VikasChaurasia *et al.,*	Simple Logic Classification	RepTree RBF Network Simple Logistic	74.47%	Weka
K.Sivakami *et al.,*	Hybrid Classification Algorithm	DT and SVM algorithms	91%	Weka

VI. CONCLUSION

Data mining is the method of defining patterns in enormous data sets comprising methods at the association of statistics, machine learning and database systems. The completion goal of the data mining process is to mine information from the dataset and change it into an understandable format for additional use.The cancer has come to be the chief source of death worldwide. The best effective way to lessen cancer deaths is to notice it formerly. Several people evade cancer screening due to the price involved in taking numerous tests for analysis. The prediction system may offer easy and a costeffective way for showing cancer and may play a key role in the previous analysis process for dissimilar types of cancer and offer effective preventive scheme. In this paper, we have discussed the importance of predicting the breast cancer in the early stage.

REFERENCES

[1] Hemlata Sahu, Shalini Shrma, and Seema Gondhalakar, "A Brief Overview on Data Mining Survey", *International Journal of Computer Technology and Electronics Engineering (IJCTEE),* Vol. 1, No.3, pp. 114-121, 2011.

[2] P. Ramachandran, N. Girija, and T. Bhuvaneswari, "Early Detection and Prevention of Cancer using Data Mining Techniques",*International Journal of Computer Applications*, Vol. 97, No. 13, pg. 48-53, July 2014.

[3] K. Arutchelvan and R. Periyasamy, "Cancer Prediction System Using Data Mining Techniques", *International Research Journal of*

Engineering and Technology, Vol.2, No. 8, pp. 1179-1183, Nov. 2015.

[4] V. Krishnaiah, G. Narsimha, and N. Subhash Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques", *International Journal of Computer Science and Information Technologies*, Vol.4, No.1, pp. 39-45, 2013.

[5] Sahar Mokhtar and Alaa. M. Elsayad, "Predicting the Severity of Breast Masses with Data Mining Methods", *International Journal of Computer ScienceIssues*, Vol. 10, Mar. 2013.

[6] Ada and Rajneet Kaur, "Using Some Data Mining Techniques to Predict the Survival Year of Lung Cancer Patient", *International Journal of Computer Science and Mobile Computing*, Vol. 2, No. 4, pp.1-6, Apr. 2013.

[7] Charles Edeki and Shardul Pandya, "Comparative Study of Data Mining and Statistical Learning Techniques for Prediction of Cancer Survivability", *Mediterranean journal of Social Sciences*., Vol. 3, No. 14, pp. 49-56, Nov. 2012.

[8] Zakaria Sulimanzubi and Rema Asheibani Saad, "Improves Treatment Programs of Lung Cancer using Data Mining Techniques", *Journal of Software Engineering and Applications*, Vol. 7, pp. 69-77, Feb. 2014.

[9] Shivangi Bhardwaj, "Data Mining Clustering Techniques – A Review", *International Journal of Computer Science and Mobile Computing*, Vol. 6, No.5, pp. 183-186, May. 2017.

[10] C. Kalaiselvi and G.M. Nasira, "A New Approach for Diagnosis of Diabetes and Prediction of Cancer using ANFIS", *World Congress on Computing and Communication Technologies*, pp. 188-190, 2014.

[11] G. Ravi Kumar, G. A. Ramachandra and K. Nagamani, "An Efficient Prediction of Breast Cancer Data using Data Mining Techniques", *International Journal of Innovations in Engineering and Technology*, Vol. 2, No. 4, pp. 139-144, Aug. 2013.

[12] Vikas Chaurasia and Saurabh Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability", *International Journal of Computer Science and Mobile Computing*, Vol. 3, No.1, pp. 10-22, Jan. 2014.

[13] K. Sivakami, "Mining Big Data: Breast Cancer Prediction using DT - SVM Hybrid Model", *International Journal of Scientific Engineering and Applied Science*, Vol. 1, No. 5, pp. 418-429, Aug. 2015.