International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 04, April 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

# A SURVEY ON PREDICTIVE ANALYSIS OF CANCER SURVIVABILITY RATE USING MACHINE LEARNING ALGORITHM

**Arjun Sharma[1], Karthik G S[2], Nidhi Mittal[3], Sindhu V L[4], Pradeep K R[5]**

[1,2,3,4]*Dept. of CSE, K.S Institute of Technology, Bengaluru, India*

[5]*Asst Professor, Dept. of CSE, K.S. Institute of Technology, Bengaluru*

## ABSTRACT

*Cancer has been described as a heterogeneous disease comprising of various subtypes. The early determination and visualization of a tumor has turned into a need in cancer research, as it can encourage the consequent clinical administration of patients. The significance of arranging patients into high or low risk groups has driven numerous examination groups, from the biomedical and the bioinformatics field, to concentrate on the utilization of machine learning (ML) strategies. Subsequently, these methods have been used as intended to demonstrate the movement and treatment of cancerous conditions. Moreover, the capacity of ML devices to distinguish key features from complex datasets uncovers their significance. An assortment of these systems, including Bayesian Networks (BNs) and Decision Trees (DTs) have been generally applied in cancer research for the advancement of prescient models, bringing about compelling and precise decision making. The prescient models talked about here depend on different managed Machine Learning systems and in addition on various input features and tests.*

***Keywords: Big data, Cancer, Data mining, Machine Learning.***

## I. INTRODUCTION

Cancer is a kind of ailment which causes the cells of the body to change its attributes and cause abnormal development of cells. Most sorts of the cancer cells in the end turn into a mass called tumor.

Breast cancer has become one of the most common diseases among women that lead to death. Breast cancer can be analyzed by grouping tumors. There are two distinct sorts of tumors, malignant and benign tumors. Doctors require a dependable analysis method to recognize these tumors. In any case, for the most part it is exceptionally hard to recognize tumors even by the specialists. Subsequently computerization of analytic framework is required for diagnosing tumors. Numerous specialists have endeavored to apply machine learning techniques for distinguishing survivability of tumors in people and it is likewise been demonstrated by the scientists that these techniques work better in diagnosing cancer.

Lung cancer starts in the tissues of the lungs. Lung cancer is the main source of tumor passing in both men and woman. Like different diseases, lung cancer happens after rehashed affront to the hereditary material of the cell. By far the most widely recognized wellspring of these put-downs is tobacco smoke, which is in charge of

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

around 85% of lung tumor deaths. The rate of lung tumor in different nations takes after their smoking examples. Some different cancer-causing agents known to bring about lung growth are found in the working environment. Lung Cancer is the uncontrolled development of strange cells; begin off in one or both lungs, for the most part in the line air sections. The abnormal cells don't form into healthy lung tissue; they give quickly and frame tumors.

## II. ANALYSIS WITH MACHINE LEARNING AND DATA MINING

Machine learning is ideal for exploiting the opportunities hidden in big data. The term "big data" regularly alludes just to the utilization of prescient analytics, user conduct analytics, or certain other propelled data analytic strategies that concentrate on values from data, and occasionally to a specific size of data set. Big data is high volume, high speed, as well as high assortment data resources that require new types of handling to enable decision making, understanding revelation and process enhancement. Big Data represents the Information resources portrayed by High Volume, Velocity and Variety to require particular Technology and Analytical Methods for its change into “Value”. [1] Additionally, another V “Veracity”[2] is added by a few associations to depict it, revisionism tested by some industry authorities.[3] The 3Vs have been extended to other reciprocal attributes of big data. [4-5] The basic characteristics are:

- Volume: The amount of produced and put away data. The extent of the data decides the esteem and potential knowledge and whether it can really be viewed as big data or not.
- Variety: The sort and nature of the data. These people groups who break down it to adequately utilize the subsequent knowledge.
- Velocity: In this unique circumstance, the speed at which the data is created and prepared to meet the requests and difficulties that lies in the way of development and improvement.
- Variability: Inconsistency of the data set can hamper procedures to handle and oversee it.
- Veracity: The quality of encapsulated information can differ considerably, influencing precise investigation.

Machine Learning, a branch of Artificial Intelligence, relates the issue of learning from data tests to the general notion of reason. [6-7] It is a scientific regulation concerned with the outline and advancement of calculations that permit PCs to develop algorithms in view of exact information, for example, from sensor information or databases. Machine Learning (ML) goes for giving computational methods to gathering, changing and updating data in artful structures, and particularly learning frameworks that will help us to actuate data from delineations. Machine learning systems are profitable in circumstances where algorithmic courses of action are not available, there is non-attendance of formal models, or the data about the application space is inadequately portrayed. ML has additionally been demonstrated an intriguing zone in bio-medical research with numerous applications, where an acceptable speculation is acquired via looking through a n-dimensional space for a given arrangement of natural specimens, utilizing distinctive systems and algorithms[8]. Learning from patient data encounters a couple challenges, since these datasets are depicted by incompleteness (missing parameter values), incorrectness (systematic or sporadic noise in the data), sparseness (few and also non-represent able patient records available), and incorrectness (wrong selection of parameters of the given errand). These algorithms are generally used for their illustration planning limits and their human like characteristics (hypothesis, energy to uproar), with a

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

particular ultimate objective to improve remedial essential administration Machine learning is fundamentally assembled into two sorts- supervised and unsupervised.

The larger part of practical machine learning utilizes supervised learning. Supervised learning is the place you have input factors (x) and a yield variable (Y) and you utilize a calculation to take in the mapping capacity from the contribution to the yield.

Unsupervised learning is the place you just have input data (X) and no comparing yield factors. The objective for unsupervised learning is to show the basic structure or dispersion in the data with a specific end goal to take in more about the data.

Dr. S. Santhoshbaboo and S. Sasikala [17] have done a review on data mining procedures for quality determination characterization. The article managed most utilized data mining procedures for quality Selection and disease grouping; especially they have concentrated on four principle developing fields. They are neural system based techniques, machine learning algorithms, hereditary techniques and Cluster based algorithms and they have indicated future change in this field.

The data mining comprises of different techniques. Diverse techniques fill different needs, every strategy offering its own points of interest and detriments. Classification and clustering are the two most regular procedures of data mining which are utilized as a part of field of medical science. Nonetheless, most data mining strategies ordinarily utilized are of supervised class as the connected expectation procedures allocate patients to either a ”benign” group that is non-dangerous or a ”malignant” group that is destructive and produce rules for the same. Consequently, the cancer disease indicative issues are fundamentally in the extent of the generally examined arrangement issues. In data mining, classification is a standout amongst the most essential tasks. It maps data into predefined targets. It is a supervised learning as targets are predefined.

The various machine learning algorithms that can be implemented for diagnosing cancer are:

**Neural systems:**

A neural system is a model that is planned by the way human nervous systems for example, minds that procedure the data. Neural systems, with their noteworthy capacity to get importance from convoluted or loose data, can be utilized to extract designs and identify patterns that are too perplexing to possibly be seen by either people or other PC strategies. Numerous neural system models, even natural neural systems accept numerous disentanglements over genuine organic neural systems.

Such simplifications are necessary to understand the intended properties and to attempt any mathematical analysis. Regardless of the possibility that every one of the properties of the neurons is known, rearrangements are still required for expository reason. Neural systems are versatile factual gadgets. This implies they can change (synaptic weights) as a component of their execution. In ANNs, every one of the neurons are working in the meantime, which makes ANN to perform undertakings at much speedier rate.[10]

ANNs handle an assortment of grouping or example acknowledgment issues. They are prepared to produce a yield as a mix between the information factors. Different concealed layers that that represent the neural connections mathematically are ordinarily utilized for this procedure. Despite the fact that ANNs serve as a highest quality level technique in a few grouping tasks [11] they experience the ill effects of specific disadvantages. Their non-specific layered structure turns out to be tedious while it can prompt exceptionally

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

poor execution. Also, this particular system is described as a "black-box". Attempting to discover how it plays out the arrangement procedure or why an ANN did not work is practically difficult to distinguish.

Tuba kiyan et al. 2004[12] has examined that factual neural systems can be utilized to perform breast cancer growth determination viably. The researcher has looked at measurable neural system with Multi Layer Perception on WBCD database. Outspread premise function(RBF), General Regression Neural Network(GRNN),Probabilistic Neural Network(PNN) were utilized for characterization and their general execution were 96.18% for Radial Basis Function (RBF), 97% PNN, 98.8% for GRNN and 95.74% for MLP. Hence it is proved that these statistical neural network structures can be applied to diagnose breast cancer.

**Support vector machine (SVM):**

A support vector machine (SVM) is an idea in insights and software engineering for an arrangement of related managed learning strategies that dissect information and perceive designs, utilized for characterization and relapse examination. The standard SVM takes an arrangement of information and predicts, for every given information, which of two conceivable classes frames the information, making the SVM a non-probabilistic binary linear classifier. However, Relevance vector machine (RVM) gives more exact results than support vector machines. This has been demonstrated by applying RVM in other disease finding, for example, ovarian growth, optical tumor and general cancer characterizations. Consequently Relevance vector machine can likewise be connected to achieve best come about for diagnosing cancer.[10]

Ilias Maglogiannis [13] et al. 2009 have displayed an article on An intelligent system for automated breast cancer growth analysis and visualization utilizing SVM based classifiers with Bayesian classifiers and ANN for prognosis and determination of breast cancer ailment. Wisconsin diagnostic breast cancer growth datasets were utilized to execute SVM model to give qualification between the malignant & benign breast masses. These datasets include estimation taken by Fine Needle Aspirates (FNA). The article gives the execution points of interest alongside the relating come about for all the evaluated classifiers. A few near studies have been completed concerning both the anticipation and determination issue showing the predominance of the proposed SVM calculation as far as affectability, specificity and precision.

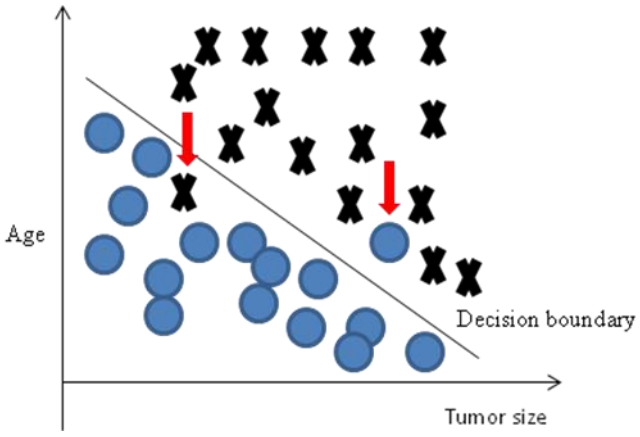


**Fig.1 An abridged model of a linear SVM classification of the input data. Figure was transcribed from the ML lectures of [18]. Tumors are classified according to their size and the patient's age. The illustrated arrows demonstrate the misclassified tumors**.

International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 04, April 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

**Naive Bayes(NB)** :

The Naive Bayes is a brisk strategy for production of statistical predictive models. NB depends on the Bayesian theorem. This supervised method examines the relationship between each property and the class for every occurrence to determine a restrictive likelihood for the connections between the property values and the class. Amid preparing, the likelihood of every class is registered by numbering how often it happens in the preparation dataset. This likelihood turns into the product of the probabilities of every single characteristic. At that point the probabilities can be evaluated from the instances in the training set.[14]

**Decision Trees (DT):**

Decision tree is a tree where each non-terminal node represents a test or decision on the considered data item. Decision of a specific branch relies on the result of test. To characterize a specific information item, we begin at root node and take after the statements down until we achieve a terminal node (or leaf). A choice is made when a terminal node is drawn closer. Decision trees moreover can be deciphered as an exceptional type of rule set, portrayed by their various leveled association of standards. Figure underneath portrays a delineation of a DT with its components and rules.[14]

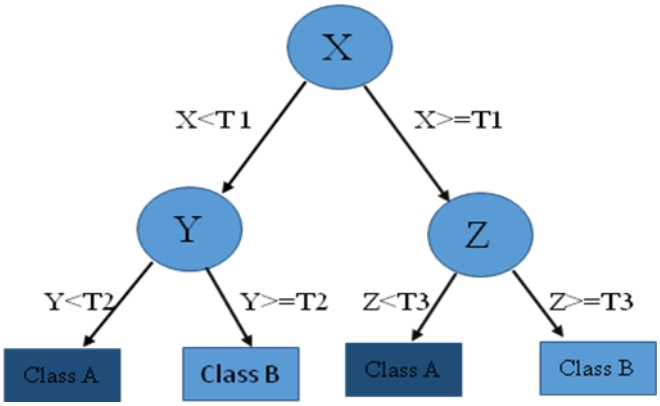


**Fig.2 An outline of a DT demonstrating the tree structure. Every variable (X, Y, Z) is depicted by a circle and the choice results by squares (Class A, Class B). T(1–3)depicts to the edges (arrangement rules) so as to effectively order every variable to a class name.**

| Decision Tree Algorithm | Support Vector Machine | Neural Network | Naive Bayes Classifier |
|---|---|---|---|
| Trees created from Numerical dataset are complex. | Support vector machines are sensitive to noisy data. | Determining the network structure and parameters is difficult | Performs well for both categorical and continuous data. |
| Best Suited for categorical data. Not suited for missing and noisy data. | It only considers 2 classes. | It is a black box model. It doesn't assume any relationship between the independent variables and let the data define the functional relationship. | Best suited for noisy data. |
| Output attribute should be categorical. | Selection of kernel function is very difficult. | High speed prediction is a great challenge in NN. | A clear dependency is Defined between attributes. |

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

| | | | |
|---|---|---|---|
| As the number of data increases, no of operations increase to build the tree model. | It suffers from over-fitting. | It also suffers from over-fitting. | It handles over-fitting. |
| They produce very complex production rules. | It takes long time to process. | It consumes lot of memory. | It consumes very less memory. |

**Table.1: Comparative study of different algorithms of Machine Learning is given above**

## III. CASE STUDY: PREDICTION OF CANCER SURVIVABILITY

About portion of all machine learning ponders on cancer prediction were centered on foreseeing persistent survivability (either 1 year or 5 year survival rates). One paper specifically compelling (Futschik et al. 2003) [15] utilized a half and half machine learning way to deal with foresee results for patients with Diffuse Large B-Cell Lymphoma (DLBCL). In particular, both clinical and genomic (microarray) information were joined to make a solitary classifier to anticipate survival of DLBCL patients. This approach contrasts to some degree from the investigation of Listgarten et al. (2004) [16] which just utilized genomic (SNP) information in its classifier pattern. Futschik et al. guessed, accurately, that clinical data could advance microarray information with the end goal that a consolidated indicator would perform superior to anything a classifier in view of either microarray information alone or clinical information alone. In gathering the test and preparing tests, the creators gathered microarray expression information and clinical data for 56 DLBCL patients. The clinical data was acquired from the International Prediction Index (IPI) which comprises of an arrangement of hazard components, that when appropriately evaluated, permits patients to be isolated into gatherings running from generally safe to high-chance. The information from the patient's IPI groupings was then used to make a straightforward Bayesian classifier. This classifier accomplished a precision of 73.2% in anticipating the mortality of DLBCL patients. The information from the patient's IPI characterizations was then used to make a basic Bayesian classifier. This classifier accomplished an exactness of 73.2% in anticipating the mortality of DLBCL patients. Independently from the Bayesian classifier, a few distinct sorts of "Evolving Fuzzy Neural Network" (EFuNN) classifiers were additionally created to handle the genomic information. The best EFuNN classifier utilized a subset of 17 qualities from the microarray information. This ideal EFuNN had a precision of 78.5%. The EFuNN classifier and the Bayesian classifier were then consolidated into a various leveled measured framework to create an accord expectation. This crossover classifier achieved a precision of 87.5%, an unmistakable change over the execution of either classifier alone. This was likewise 10% superior to anything the best performing machine learning classifier (77.6% by SVMs). The EFuNN classifier was approved utilizing a one cross-approval methodology. This was likely because of the little specimen estimate. No outside approval set was accessible to test the all inclusive statement of the model. With just 56 patients (tests) being ordered by means of 17 quality components, the sample per feature ratio (SFR) is a little more than 3. When in doubt, a SFR of fewer than 5 does not really ensure a powerful classifier (Somorjai et al. 2003). In any case, it is very clear that the creators knew about this issue and went to impressive lengths to legitimize their approach by clarifying, in detail, the internal workings of their classifier. This incorporated a portrayal of how the Bayesian classifier was constructed, how the EFuNN works, and how the two classifiers cooperate to give a solitary forecast.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Furthermore, the creators additionally examined, and therefore affirmed, the freedom of the microarray information from the clinical information. This meticulousness is especially excellent for a machine learning examination of this kind. This concentrate pleasantly shows how the force of utilizing both clinical and genomic information in cancer visualization can significantly upgrade forecast precision.

## IV. CONCLUSION

In this survey, we see the various machine learning and data mining techniques that can be applied to predict the survivability rates in cancer patients. The classification algorithms are used to identify the various categories of breast cancer and lung cancer. Here, an attempt is made to atomize the process of predicting the survivability rate based on the input data set.

## V. ACKNOWLEDGEMENT

Authors would like to thank VGST (Vision Group on Science and Technology), Government of Karnataka, India for providing infrastructure facilities through the K-FIST Level I project at KSIT, CSE R&D Department, Bengaluru.

## REFERENCES

[1] De Mauro, Andrea; Greco, Marco; Grimaldi, Michele (2016). "A Formal definition of Big Data based on its essential Features". Library Review. **65**:122–135. doi:10.1108/LR-06-2015-0061.

[2] Seth. "Big Data: Avoid 'Wanna V' Confusion". InformationWeek, Retrieved 5 January.

[3] New Horizons for a Data-Driven Economy – Springer. Doi:10.1007/978-3-319-21569-3.

[4] Hilbert, Martin. "Big Data for Development: A Review of Promises and Challenges. Development Policy Review.". Martinhilbert.net. Retrieved 2015-10-07.

[5] DT&SC 7-3: What is Big Data?12 August 2015 – via youtube.

[6] Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.

[7] Witten IH, Frank E. Data mining: practical machine learning tools and techniques.

[8] Niknejad A, Petrovic D. Introduction to computational intelligence techniques and areas of their applications in medicine. Med Appl Artif Intell 2013;51.

[9] Dr.Santhosh baboo, S.Sasikala "A Survey on data mining techniques in gene selection and cancer classification"-April 2010 International journal of Computer science and information technology.

[10] Breast cancer diagnosis using machine learning algorithms –a survey By B.M.Gayathri ,C.P.Sumathi and T.Santhanam, International Journal of Distributed and Parallel Systems (IJDPS) Vol.4, No.3, May 2013.

[11] Ayer T, Alagoz O, Chhatwal J, Shavlik JW, Kahn CE, Burnside ES. Breast cancerrisk estimation with artificial neural networks revisited. Cancer 2010;116:3310–21.

[12] Chih-Lin Chi, W. Nick Street, and William H. Wolberg " Application of Artificial Neural Network-Based Survival Analysis on Two Breast Cancer Datasets"- AMIA Annu Symp Proc. 2007; 2007:130–134.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

[13] Ilias Maglogiannis, E Zafiropoulos "An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers" Applied Intelligence, 2009 – Springer.

[14] A comparative survey on data mining techniques for breast cancer diagnosis and prediction-Survey By Hamid Karim KhaniZand, Indian Journal of Fundamental and Applied Life Sciences ISSN: 2231– 6345, 2015 Vol.5 (S1), pp. 4330-4339.

[15] Futschik ME, Sullivan M, Reeve A, et al. 2003. Prediction of clinical behaviour and treatment for cancers. Appl Bioinformatics, 2(3 Suppl): S53-8.

[16] Listgarten J, Damaraju S, Poulin B et al. 2004. Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. Clin Cancer Res, 10:2725-37.

[17] Dr.Santhoshbaboo, S.Sasikala "A Survey on data mining techniques in gene selection and cancer classification"-April 2010 International journal of Computer science and information technology.

[18] Adams S. Is Coursera the beginning of the end for traditional higher education?Higher Education; 2012.