# Machine Learning Techniques Used for Cancer Disease – A Review

**Ghantasala Venu Gopal**
Research scholar, Department of Computer science, Rayalaseema University.
**Dr. R. B. V. Subramanyam**
Research guide, Professor, Department of Computer science, NIT, Warangal

**Abstract:**

Cancer is a disease caused by uncontrolled division of abnormal cells in a part of the body. There are various types of Cancers in the world. In earlier days, the diagnosis of cancer mainly depended on the Doctor's experience and knowledge. But from the last decade, decision support system with computers has been playing a vital role in health care industry. For early prevention and detection of the cancer patients, Machine Learning techniques are used. The rapid growth of the Machine Learning Techniques truly helped the Doctors to take appropriate decisions in diagnosis. In this paper, the commonly used Machine Learning Techniques for the Cancer prediction are summarized.

**Keywords**: Machine Learning, Supervised Learning, Support Vector Machines and Neural Networks.

## I. Introduction

Cancer is one of the most common diseases in the world that results in mainstream of death caused by unrestrained growth of cells in any of the tissues or parts of the body. Cancer is that deadly disease which is caused due to change in normal cells of the body and as a result there is uncontrolled growth of cells which give rise to tumor, except leukemia this is the main cause of cancer. If tumor is not treated in time it grows and spread into surrounding areas through bloodstream and affects the digestive, resource and circulatory system and cause severe health consequences which are the important cause of death. Men are more prone to Lung, prostate, stomach, liver cancer. While women are more prone to breast, colorectal, lung, cervix uteri, and stomach cancer. Collection of related disease is called cancer. If the proper treatment and diagnosis of the disease is not done in time as it is found in most of the cases this malignant disease can even cause death.

According to WHO (World Health Organization) 8.2 million people die each year from cancer and it is estimated that 13% of total death worldwide in caused due to cancer 70% increase in new cases of cancer is expected over the next two decades. Over all 100 types of cancer exists each requiring unique treatment and diagnosis. The most commonly diagnosed cancer worldwide is of the Lungs (1.8 million, 13% of total) Breast cancer (1.7 million, 12% of total) Colorectal (1.4 million, 9.7 of total). The most common causes of cancer death are cancer of lungs (1.6 million, 19.1% of total) Liver cancer (0.8 million, 9.1% of total) Stomach cancer (0.7 million, 8.8 of total). It is estimated that by year 2025 increase to 19.3 million new cancer cases per year can be noticed due to growth and aging of the population growth. Cancer is one of the most important reasons of mortality in different countries of the world [1]. The Figure 1 shows [2] the estimated figure of total cancer cases in India which shows the seriousness of the issue.

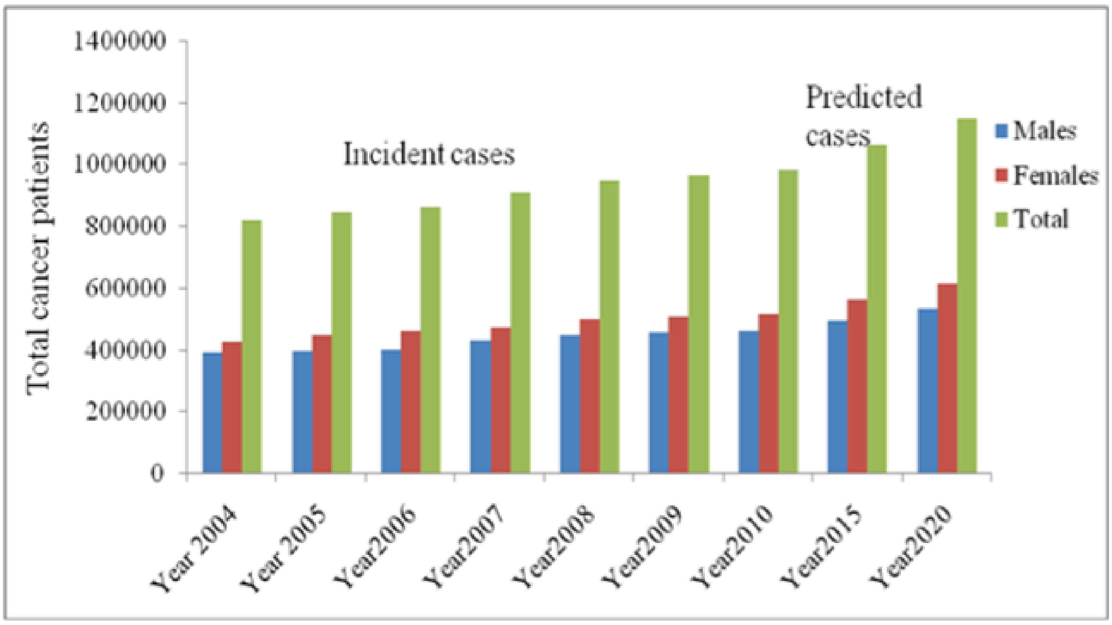

**Figure 1: Cancer Statistics Scenario in India**

This content of the paper proceeds with a brief introduction in Section I. Section II provides a detailed survey of various researches about several Machine Learning Techniques and methodologies that have been applied in reference to diagnosis and prediction of cancer disease. Section III describes various types of Machine Learning Algorithms used in medical sector. Section IV explains that Machine Learning Algorithms are often grouped by similarity in terms of their functionalities. Section V provides various applications of Machine Learning Algorithms in Medical sector. Conclusion of this analysis paper is in Section V. This paper ends with required references.

## II. Related Work

Plenty of work and researches have been done to find out different methods of diagnosis of various cancers types. It is an attempt to predict and diagnose the cancer disease based on symptoms that occurs at an early stage.

This paper [3] discusses the lung cancer which is one of the deadly diseases of lungs. Based on that, feature selection process short lists 20 such parameters. Some of those influencing parameters are weight loss, bloody mucus, back pain etc. Here, researchers focus on pre-diagnosis which is considered as the most vital stage to know the susceptible patients for going through special diagnosis process. They noticed that supervised learning ways are far better than to the cross validation approach. From this research, it was found that random tree classifier, KNN, logistic, multilayer perception, sequential minimal, optimization has given much more better and reliable performance in this domain.

This research in [4] dealt with breast cancer and the prediction of the disease was done through Artificial Neural Network (ANN), Logistic regression, Naive Bayes techniques. The objective of the research aims at giving the following outcomes; firstly, it evaluates medical data set in terms of quality grammatically and secondly, it evaluates data mining methods with respect to their applicability to the data. Finally, the knowledge extracted from the data set is used for disease prediction by applying Artificial Neural Network (ANN), Logistic Regression, and Naïve Bayes. It is found that these techniques had highest lifting factor for most of class values.

Classification based pattern analysis techniques are used in this work for diagnosing the cancer. Several well known classification algorithm such as DT (Decision Tree), SVM (Support Vector Machine), KNN (K-Nearest Neighbor) and NN (Neural Networks) are used for diagnoses of cancer. It is established that the process of classification depends on the value of various features in the collected data. Here, the authors of [5] found that medical disease data often have some noise data as well as boundary value data. They suggested techniques to deal with such noisy data. For optimization of accuracy they employed Ant Colony Optimization technique.

Authors in [6] their work used several data mining approaches like classification, classification rule mining, soft-computing techniques, Neural Networks and Fuzzy logic for diagnosis of oral cancer. They specified the effectiveness of each of the above techniques for the classification task in medical domain. Apart from it, they showed the importance of genetic algorithms in optimizing the data mining algorithm in terms of accuracy for prediction.

In Mukti & Ahmed (2013) the authors explored the applicability of Apriori and Decision Tree to discover significant discover patterns. The goal was using significant patterns to develop a lung cancer prediction system. The prediction system was able to detect a person's predisposition for lung cancer, 400 cancer and non cancer patients data were collected and evaluated. With the proposed methodology, the author showed that it is possible to find statistically significant associations from the gathered data set. However the result evaluated through the proposed methodology does not show a high degree of statistical confidence.

## III. Machine Learning

Machine Learning (ML) can be explained as automating and improving the learning process of computers based on their experiences without being actually programmed i.e. without any human assistance. The process starts with feeding good quality data and then training our machines (computers) by building machine learning models using the data and different algorithms. The choice of algorithms depends on what type of data do we have and what kind of task we are trying to automate. There are three main learning styles or learning models that an algorithm follows.

i). Supervised Learning:

Input data is called training data and has a known label or result such as spam/not-spam or a stock price at a time. A model is prepared through a training process in which it is required to make predictions and is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy on the training data. Example problems are classification and regression. Example algorithms include: Logistic Regression and the Back Propagation Neural Network.

ii). Unsupervised Learning:

Input data is not labeled and does not have a known result. A model is prepared by deducing structures present in the input data. This may be to extract general rules. It may be through a mathematical process to systematically reduce redundancy, or it may be to organize data by similarity. Example problems are clustering, dimensionality reduction and association rule learning. Example algorithms include: the Apriori algorithm and K-Means.

iii). Semi-Supervised Learning:

Input data is a mixture of labeled and unlabelled examples. There is a desired prediction problem but the model must learn the structures to organize the data as well as make predictions. Example problems are classification and regression. Example algorithms are extensions to other flexible methods that make assumptions about how to model the unlabeled data.

### IV. Machine Learning Algorithmsgrouped By Similarity

The following Machine learning algorithms/methods were used in health care. They are grouped by similarity.

**a).Regression Algorithms:** Regression is concerned with modeling the relationship between variables that is iteratively refined using a measure of error in the predictions made by the model. Regression methods are a workhorse of statistics and have been co-opted into statistical machine learning.

Example:

- Ordinary Least Squares
- Regression (OLSR)
- Linear Regression
- Logistic Regression
- Stepwise Regression

**b). Instance-based Algorithms:** Instance-based learning model is a decision problem with instances or examples of training data that are deemed important or required to the model.

Example:

- k-Nearest Neighbor (KNN)
- Learning Vector Quantization (LVQ)
- Self-Organizing Map (SOM)
- Locally Weighted Learning (LWL)
- Support Vector Machines (SVM )

**c). Regularization Algorithms:** An extension made to another method (typically regression methods) that penalizes models based on their complexity, favoring simpler models that are also better at generalizing.

- Ridge Regression
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Elastic Net
- Least-Angle Regression (LARS)

**d). Decision Tree Algorithms:** Decisions fork in tree structures until a prediction decision is made for a given record. Decision trees are trained on data for classification and regression problems. Decision trees are often fast and accurate and a big favorite in machine learning.

Example:

- Classification and Regression Tree (CART)
- Iterative Dichotomiser 3 (ID3)
- C4.5 and C5.0 (different versions of a powerful approach)
- Chi-squared Automatic Interaction Detection (CHAID)
- Decision Stump
- M5
- Conditional Decision Trees

**e). Bayesian Algorithms:** Bayesian methods are those that explicitly apply Bayes' Theorem for problems such as classification and regression.

- Naive Bayes
- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Averaged One-Dependence Estimators (AODE)
- Bayesian Belief Network (BBN)

Vol 9 Issue 04 2020
ISSN NO: 2230-5807

- Bayesian Network (BN)

**f). Clustering Algorithms:** Clustering methods are typically organized by the modeling approaches such as centroid-based and hierarchal. All methods are concerned with using the inherent structures in the data to best organize the data into groups of maximum commonality.

- k-Means
- k-Medians
- Expectation Maximization (EM)
- Hierarchical Clustering

**g).** Artificial Neural Network Algorithms: Artificial Neural Networks are models that are inspired by the structure and/or function of biological neural networks. They are a class of pattern matching that are commonly used for regression and classification problems but are really an enormous subfield comprised of hundreds of algorithms and variations for all manner of problem types.

Example:

- Perceptron
- Multilayer Perceptrons (MLP)
- Back-Propagation
- Stochastic Gradient Descent
- Hopfield Network
- Radial Basis Function Network (RBFN)

**h). Dimensionality Reduction Algorithms:** Clustering methods, dimensionality reduction seek and exploit the inherent structure in the data, but in this case in an unsupervised manner or order to summarize or describe data using less information. This can be useful to visualize dimensional data or to simplify data which can then be used in a supervised learning method.

**Example:**

- Principal Component Analysis (PCA)
- Principal Component Regression (PCR)
- Partial Least Squares Regression (PLSR)
- Multidimensional Scaling (MDS)
- Linear Discriminant Analysis (LDA)

**i). Ensemble Algorithms:** Ensemble methods are models composed of multiple weaker models that are independently trained and whose predictions are combined in some way to make the overall prediction.

Example:

- Boosting
- Bootstrapped Aggregation (Bagging)
- AdaBoost
- Stacked Generalization (Stacking)
- Gradient Boosting Machines (GBM)
- Gradient Boosted Regression Trees (GBRT)
- Random Forest

### V. Machine Learning Applications in Health Care

To take the accurate decisions in health care the Machine Learning is very needful. The following are the different applications

✓ Machine Learning provides support for constructing a model for managing the hospital resources which is an important task in healthcare. Using Machine Learning, it is possible to detect the chronic disease and based on the complication of the patient disease prioritize the patients so that they will get effective treatment in timely and accurate manner.

- ✓ Different Machine Learning approaches are used to analyze the various hospital details in order to determine their ranks. Ranking of the hospitals are done on the basis of their capability to handle the high risk patients.
- ✓ Machine Learning helps the healthcare institutes to understand the needs, preferences, behavior, patterns and quality of their customer in order to make better relation with them.
- ✓ A system for inspection is constructed using Machine Learning techniques to discover unknown or irregular patterns in the infection control data. Association rules are used to produce unexpected and interesting information from the public surveillance and hospital control data.
- ✓ Healthcare insurer develops a model to detect the fraud and abuse in the medical claims using Machine Learning Techniques. This model is helpful for identifying the improper prescriptions, irregular or fake patterns in medical claims made by physicians, patients, hospitals etc.
- ✓ American Health ways system constructs a predictive model using Machine Learning techniquesto recognize the patients having high risk. The main concern of this system is to handle the diabetic patients, improve their health quality and also offers cost savings services to the patient. Using Predictive model, healthcare provider recognize the patient which require more concern as compare to other patients
- ✓ Machine Learning play an important role for making effective policy of healthcare in order to improve the health quality as well as reducing the cost for health services.

## VI. Conclusion

The prediction of Cancer Dieses is very important aspect in Health care sector. This paper, describes the various Machine Learning Techniques, which are used to predict Cancer Diseases. It concludes that there is no single Machine Learning Technique which gives consistent results for all type of Cancer Diseases. The performance of the Machine Learning Techniques depends on the type of data set that is used in medical diagnosis. The main idea of this survey is using different Machine Learning Techniques on Cancer Diseases yields different results. These comparing results give best algorithm for future work.

## VII. References

[1]. WHO. http://www.who.int/mediacentre/factsheets/fs297/en/ Retrieved on May 20, (2016).

[2]. Satyam Shukla, Dharmendra Lal Guptaand Bakshi Rohit Prasad, “ Comparative Study of Recent Trends on Cancer Disease Prediction using Data Mining Techniques” ,International Journal of Database Theory and Application Vol.9, No.9 (2016), pp.107-118.

[3]. K. Balachandran and R. Anitha, “Ensemble based optimal classification model for pre-diagnosis of lung cancer”, 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), IEEE, (2013).

[4]. K. Shiny, “Implementation of Data Mining Algorithm to Analysis Breast Cancer”, International Journal for Innovative Research in Science and Technology, vol. 1, no. 9, (2015), pp. 207-212.

[5]. S. S. Shrivastava, V. K. Choubey and A. Sant, “Classification Based Pattern Analysis on the Medical Data in Health Care Environment”, International Journal of Scientific Research in Science, Engineering and Technology, vol. 2, no. 1, (2016).

[6] R. Vidhu and S. Kiruthika, “A New Feature Selection Method for Oral Cancer Using Data Mining Techniques”, International Journal of Advanced Research in Computer and Communication Engineering, vol. 5, no. 1, (2016).

[7].D. Tomar and S. Agarwal, “A survey on Data Mining approaches for Healthcare”, International Journal of Bio-Science and Bio-Technology, vol. 5, no. 5, (2013), pp. 241-266.

[8]. H. Li, G. Hong and Z. Guo, “Reversal DNA methylation patterns for cancer diagnosis”, 2014 8th International Conference on Systems Biology (ISB), IEEE, (2014).

[9]. R. Chau, “Determining the familial risk distribution of colorectal cancer: a data mining approach”, Familial cancer, (2015), pp. 1-11.

[10]. N. Rathore, D. Tomar and S. Agarwal, “Predicting the survivability of breast cancer patients using ensemble approach”, 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), IEEE, (2014).

[11].Berman AT, James SS, Rengan R. Structure, mechanism, and evolution of the mRNA capping apparatus. Cancers (Basel). 2015;7(3):1178–90. [3] D. Tomar and S. Agarwal, “Hybrid feature selection based weighted least squares twin support vector machine approach for diagnosing breast cancer, hepatitis, and diabetes”, Advances in Artificial Neural Systems, vol. 2015, no. 1.

[12] B. R. Prasad and S. Agarwal, “Modeling risk prediction of diabetes-A preventive measure”, 9th International Conference on Industrial and Information Systems (ICIIS), Gwalior, IEEE, (2014).pp.1-6.

[13] D. Tomar, B. R. Prasad and S. Agarwal, “An efficient Parkinson disease diagnosis system based on Least Squares Twin Support Vector Machine and Particle Swarm Optimization”, 9th International Conference on Industrial and Information Systems (ICIIS), Gwalior, IEEE, (2014), pp. 1-6.

[14] A. K. Yadav, D. Tomar and S. Agarwal, “Clustering of lung cancer data using Foggy K-means”, 2013 International Conference on Recent Trends in Information Technology (ICRTIT), IEEE, (2013).

[15].Stachnik A, Yuen T, Iqbal J, Sgobba M, Gupta Y, Lu P, et al. Repurposing of bisphosphonates for the prevention and therapy of nonsmall cell lung and breast cancer. Proc Natl Acad Sci U S A. 2014;111(50):17995–8000.

[16].Chen H, Zhang H, Zhang Z, Cao Y, Tang W. Network-based inference methods for drug repositioning. Comput Math Methods Med. 2015;2015:130620.

[17]. Lee HS, Bae T, Lee JH, Kim DG, Oh YS, Jang Y, et al. Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug. BMC Syst Biol. 2012;6:80.

[18].HuangCH,Wu MY,Chang PM, Huang CY, Ng KL.Insilico identificationofpotential targetsanddrugsfornon-smallcelllung cancer. IETSystBiol.2014;8(2):56–66.

[19]. Huang CH, Chang PM, Lin YJ, Wang CH, Huang CY, Ng KL. Drug repositioning discovery for early- and late-stage non-small-cell lung cancer. Biomed Res Int. 2014;2014:193817.

[20].Huang CH, Peng HS, Ng KL. Prediction of cancer proteins by integrating protein interaction, domain frequency, and domain interaction data using machine learning algorithms. Biomed Res Int. 2015;2015:312047.