# Cancer Type Prediction

Swapnil Singh
Machine Learning Intern
AI Technology and System
singhswap1999@gmail.com
www.ai-techsystems.com

***Abstract*— This paper is about predicting the type of the cancer. Cancer is a group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body. These contrast with benign tumors, which do not spread. Possible signs and symptoms include a lump, abnormal bleeding, prolonged cough, unexplained weight loss, and a change in bowel movements. While these symptoms may indicate cancer, they can also have other causes. Over 100 types of cancers affect humans. The risk of cancer increases significantly with age, and many cancers occur more commonly in developed countries. Rates are increasing as more people live to an old age and as lifestyle changes occur in the developing world. The financial costs of cancer were estimated at $1.16 trillion USD per year as of 2010.**

**Keyword- Machine Learning, Prediction, SVM.**

## *INTRODUCTION*

Over the past decades, a continuous evolution related to cancer research has been performed [1]. Scientists applied different methods, such as screening in early stage, in order to find types of cancer before they cause symptoms. Moreover, they have developed new strategies for the early prediction of cancer treatment outcome. With the advent of new technologies in the field of medicine, large amounts of cancer data have been collected and are available to the medical research community. However, the accurate prediction of a disease outcome is one of the most interesting and challenging tasks for physicians. As a result, ML methods have become a popular tool for medical researchers. These techniques can discover and identify patterns and relationships between them, from complex datasets, while they are able to effectively predict future outcomes of a cancer type.

Given the significance of personalized medicine and the growing trend on the application of ML techniques, we here present a review of studies that make use of these methods regarding the cancer prediction and prognosis. In these studies prognostic and predictive features are considered which may be independent of a certain treatment or are integrated in order to guide therapy for cancer patients, respectively [2]. In addition, we discuss the types of ML methods being used, the types of data they integrate, the overall performance of each proposed scheme while we also discuss their pros and cons.

An obvious trend in the proposed works includes the integration of mixed data, such as clinical and genomic. However, a common problem that we noticed in several works is the lack of external validation or testing regarding the predictive performance of their models. It is clear that the application of ML methods could improve the accuracy of cancer susceptibility, recurrence and survival prediction. Based on [3], the accuracy of cancer prediction outcome has significantly improved by 15%–20% the last years, with the application of ML techniques.

Several studies have been reported in the literature and are based on different strategies that could enable the early cancer diagnosis and prognosis. Specifically, these studies describe approaches related to the profiling of circulating mi RNAs that have been proven a promising class for cancer detection and identification. However, these methods suffer from low sensitivity regarding their use in screening at early stages and their difficulty to discriminate benign from malignant tumours. Various aspects regarding the prediction of cancer outcome based on gene expression signatures are discussed in These studies list the potential as well as the limitations of microarrays for the prediction of cancer outcome. Even though gene signatures could significantly improve our ability for prognosis in cancer patients, poor progress has been made for their application in the clinics. However, before gene expression profiling can be used in clinical practice, studies with larger data samples and more adequate validation are needed.

In the present work only studies that employed ML techniques for modelling cancer diagnosis and prognosis are presented.

## *RELATED WORK*

AlirezaOsarech, BitaShadgar used SVM classification technique on two different benchmark datasets for breast cancer which got 98.80% and 96.63% accuracies. MandeepRana, PoojaChandorkar, AlishibaDsouza worked on the diagnosis and the prediction of recurrence of breast cancer by applying KNN, SVM, Naïve Bayes and Logistic Regression techniques, programmed in MATLAB. The classification techniques are applied on two datasets taken from UCI depository. A dataset of them is used for identification of disease(WDBC) and the next one is used for recurrence prediction (WPBC)[3].VikasChaurasia, BB Tiwari and Saurabh Pal used three famous algorithms such as J48, Naive bayes, RBF, to build predictive models on breast cancer prediction and compared their accuracy. The results had shown that Naive Bayes predicted well among them with an accuracyof97.36% [4]. Haifeng Wang and Sang Won Yoon compared Naive Bayes Classifier, Support Vector Machine (SVM), AdaBoost tree, Artificial Neural Networks (ANN), to find a powerful model for breast cancer prediction. They Prediction of Breast Cancer Using

Supervised Machine Learning Techniques 1107 Retrieval Number: F3384048619/19©BEIESP Published By: Blue Eyes Intelligence Engineering & Sciences Publication

Implemented PCA for dimensionality reduction. I worked on breast cancer prediction and stated that artificial neural networks are widely used. The paper featured about the advantages and short comings of using machine learning methods like SVM, Naive Bayes, Neural network and Decision trees. I took data from Wisconsin Breast Cancer database and worked on breast cancer diagnosis..The results of their experiments proved that ANN and Logistic Algorithm worked better and provided a good solution. It achieved an accuracy of 98.50%

## *METHODOLOGIES*

### *DIMENSIONALITY REDUCTION*

Dimensionality Reduction is a process in which the number of independent variables is reduced to a set of principle variables by removing those which are less significant in predicting the outcome. Dimensionality Reduction is used to get two dimensional data so that better visualization of machine learning models can be done by plotting the prediction regions and the prediction boundary for each model. Whatever may be the number of independent variables, we often end up with two independent variables by applying a suitable dimensionality reduction technique. There are two methods, namely Feature selection and Feature Extraction.

### *FEATURE SELECTION*

Feature selection is finding the subset of original features by different approaches based on the information they provide, accuracy, prediction errors.

### *FEATURE PROJECTION*

Feature projection is transformation of high-dimensional space data to a lower dimensional space (with few attributes). Both linear and nonlinear reduction techniques can be used in accordance with the type of relationships among the features in the dataset. The dataset used in this research is a multidimensional dataset with 32 attributes, which are related to cell parameters. Selection of features by the application of feature selection is a complex task. Moreover, it cannot give the most accurate features. Therefore we have applied a feature projection technique, PCA to derive two principal components from the dataset.

### *MODEL SELECTION*

The most exciting phase in building any machine learning model is selection of algorithm. We can use more than one kind of data mining techniques to large datasets. But, at high level all those different algorithms can be classified in two groups: supervised learning and unsupervised learning. Supervised learning is the method in which the machine is trained on the data which the input and output are well labeled. The model can learn on the training data and can process the future data to predict outcome. They are grouped to Regression and Classification techniques. A regression problem is when the result is a real or continuous value, such as “salary” or “weight”. A classification problem is when the result is a category like filtering emails “spam” or “not spam”. Unsupervised Learning: Unsupervised learning is giving away information to the machine that is neither classified nor labelled and allowing the algorithm to analyse the given information without providing any directions. In unsupervised learning algorithm the machine is trained from the data which is not labelled or classified making the algorithm to work without proper instructions. In our dataset we have the outcome variable or Dependent variable i.e. Y having only two set of values, either M (Malign) or B(Benign). So Classification algorithm of supervised learning is applied on it. We have chosen three different types of classification algorithms in Machine Learning.

1– Logistic Regression
2– Support Vector Machine

### *1 - Logistic Regression*

Logistic Regression is a supervised machine learning technique, employed in classification jobs ( for predictions based on training data).Logistic Regression uses an equation similar to Linear Regression but the outcome of logistic regression is a categorical variable whereas it is a value for other regression models. Binary outcomes can be predicted from the independent variables. The outcome of dependent variable is discrete. Logistic Regression uses a simple equation which shows the linear relation between the independent variables. These independent variables along with their coefficients are united linearly to form a linear equation that is used to predict the output. The equation used by basic logistic model is

$$\text{Ln}\left(\frac{\pi}{1-\pi}\right)=a0+a1*x+a2*x \qquad (1)$$

This is called the logistic function This algorithm is entitled as logistic regression as the key method behind it is logistic function. The output can be predicted from the independent variables, which form a linear equation. The output predicted has no restrictions, it can be any value from negative infinity to positive infinity. But the output required is a class variable (i.e., yes or no, 1 or 0). So, the outcome of the linear equation should be flattened into a small range (i.e [0,1]).Logistic function is used here to suppress the outcome value between 0 and 1. Logistic function can also be called sigmoid function or Cost function. Logistic function is a Shaped curve which takes the input (numeric value) and changes it to a value between 0 and 1. Applying antilog on both sides of the above equation gives the eq(2) in

$$y=\frac{e^{a0+a1*x+a2x2}}{1+e^{a0+a1*x+a2x2}} \qquad (2)$$

which the predicted value is y and a0 is the y intercept and a1, the coefficient of the independent variable x1(principal component) a2 is the coefficient of the independent variable x2 and e is the base of natural algorithm. In our research the principal components (pc1 and pc2) derived from the dimensionality reduction replace the independent variables x1 and x2. The y intercept and the regression coefficients are

estimated by the maximum likelihood estimation method rather than least squares method of estimation.

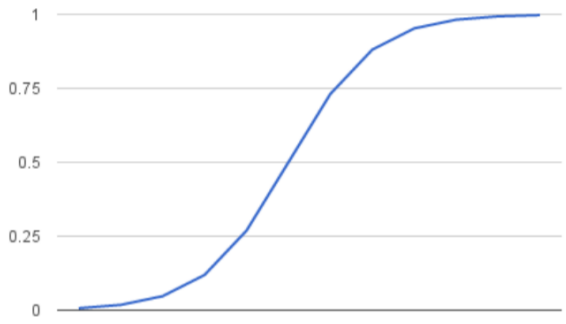


**Fig 1.**Logistic Function

***SUPPORT VECTOR MACHINE***

Support Vector Machine is a supervised machine learning algorithm which is doing well in pattern recognition problems and it is used as a training algorithm for studying classification and regression rules from data. SVM is most precisely used when the number of features and number of instances are high. A binary classifier is built by the SVM algorithm [13]. This binary classifier is constructed using a hyper plane where it is a line in more than 3-dimensions.The hyper plane does the work of separating the members into one of the two classes.

Hyper plane of SVM is built on mathematical equations. The equation of hyper plane is WTX=0 which is similar to the line equation y= ax + b. Here W and X represent vectors where the vector W is always normal to the hyper plane. WTX represents the dot product of vectors. As SVM deals with the dataset when the number of features are more so, we need to use the equation WTX=0 in this case instead of using the line equation y= ax + b. If a set of training data is given to the machine, each data item will be assigned to one or the other categorical variables, a SVM training algorithm builds a model that plots new data item to one or the other category. In an SVM model, each data item is represented as points in an n-dimensional space where n is the number of features where each feature is represented as the value of a particular coordinate in the n-dimensional space. Classification is carried out by finding a hyper-plane that divides the two classes proficiently. Later, new data item is mapped into the same space and its category is predicted based on the side of the hyper-plane they turn up

## *CONCLUSION*

Our work mainly focused in the advancement of predictive models to achieve good accuracy in predicting valid disease outcomes using supervised machine learning methods. The analysis of the results signify that the integration of multidimensional data along with different classification, feature selection and dimensionality reduction techniques can provide auspicious tools for inference in this domain.

Further research in this field should be carried out for the better performance of the classification techniques so that it can predict on more variables.

## *REFERENCE*

[1]. Data Set Source : https://doi.org/10.17632/sf5n64hydt.1

[2]. Yi-Sheng Sun, Zhao Zhao, Han-Ping-Zhu,"Risk factors and Preventions of Breast Cancer" International Journal of Biological Sciences.

[3]. AlirezaOsarech, BitaShadgar,"A Computer Aided Diagnosis System for Breast Cancer",International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011

[4]. MandeepRana, PoojaChandorkar, AlishibaDsouza, "Breast cancer diagnosis and recurrence prediction using machine learning techniques", International Journal of Research in Engineering and Technology Volume 04, Issue 04, April 2015.

[5]. VikasChaurasia, BB Tiwari and Saurabh Pal – "Prediction of benign and malignant breast cancer using data miningstechniques",Journal of Algorithms and Computational Technology

[6]. Haifeng Wang and Sang Won Yoon – Breast Cancer Prediction using Data Mining Method, IEEE Conference paper

[7]. D.Dubey ,S.Kharya, S.Soni and –"Predictive Machine Learning techniques for Breast Cancer Detection", International Journal of Computer Science and Information Technologies,Vol.4(6),2013,1023-1028.