International Journal of Engineering, Science and Mathematics
Vol. 8 Issue 4, April 2019,
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijesm.co.in, Email: ijesmj@gmail.com

IDENTIFY BREAST CANCER USING MACHINE LEARNING ALGORITHM

A.Mohana Priya*
Dr.K.E.Kannammal**
S.P.Kavya***
R.Iswarya****
M.Nivetha*****

ABSTRACT

Breast cancer is one of the severe disease among women worldwide. In this paper Machine Learning algorithm is used to predict the severity of breast cancer. Clinical Dataset is utilized for analysis using Machine Learning algorithm called Support Vector Machine. In this area cancer growth also reviewed stage by stage and suitable diagnosis can be suggested to prevent them from its severity. As an extension to the previous researches related to the discovery of breast cancer, we proposed a model to help in resolving the difficulty of determining the degree of risk for the disease and to get best practices, abatement time and expense with the objective of advancing wellbeing, based on data collected from hospitals. The model is applying classification techniques such as Support vector machine on the collected breast cancer data, which in turn predicts the severity of breast cancer. After evaluation and testing using the mentioned classification techniques on the breast cancer dataset, we obtained an accuracy of 94%, which is an accepted rate of prediction for the severity of breast cancer. This paper will discuss about the accuracy of the algorithm and prediction using dataset.

KEYWORDS:

Support Vector Machine;
Clinical Dataset;
Breast Cancer;
Classification.

Author correspondence:

A. Mohana Priya,
Assistant Professor,
Sri Shakthi Institute of Engineering and Technology,

1. INTRODUCTION

Cancer is the word given to a collection of relevant diseases. In all types of cancer, some of the body’s cells begin to propagate without stopping and spread into surrounding tissues [1]. Cancer can almost starts anywhere in the human body, which is made up of trillions of collections of cells. Normally, human cells grow more and more and divide to form new cells as the body needs them for its functioning. When cells grow old or become damaged, they die, and new cells take their place [2]. When cancer develops, however, this orderly process breaks down. As cells become more abnormal, old ordamaged cells survive when they should die, and new cells are form when they are not needed. These extra cells can be propagated without stopping and may form growths called tumors. In another meaning cancer is one type of disease. It happens when cells growth in a part of the human body becomes out-of control. In other words, whenever cells in part of the body divide uncontrollably and damage the other cells, cancer has occurred.

Nowadays, more than 100 types of cancers based on the part of body where it’s appeared, or cells that are affected, have been classified. Currently, cancer has become one of the main causes of death all over the world. Several factors affect the creation or spreading cancers including: gender, age, genetics, marital status, quality of life, living location etc. [3]. In the traditional model for transforming data to knowledge, some manual analysis and interpretation are executed. For example, in medical centers, generally doctors or specialists manually analyze current trends, disease and health-care data, then make a report and use this report for decision making or planning for medical diagnosis, treatments etc. [9]. The problem of this type of data analysis is that, this form of manual data analysis is slow, expensive, time consuming, and highly subjective.

In recent years machine learning has become widely used in many areas for the classification or prediction of the cancer disease especially breast cancer. There are various major machine learning algorithm that have been developed and used.Projects recently for knowledge discovery from databases which have the information about the patients.

In this paper we explain that how this study will support the doctors for predicting a patient's severity condition by applying the proposed intelligence models.

2. COMPARISION STUDY

It is very important that which algorithm should be selected to predict the results from dataset. Most popular algorithms which are proposed for cancer prediction is listed below.

- Random Forest Algorithm
- Artificial Neural Network
- Support Vector Machine

a)Random Forest Algorithm(RF):

Random Forest is a flexible, easy to use machine learning algorithm that can produces, even without hyper-parameters tuning, a great result gained most of the time. It is also one of the most used algorithms, because it's simplicity and the fact that it can be used for both classification and regression tasks.

Random Forest is a supervised learning algorithm. It creates a forest of decisions and makes it somehow random. The forest it builds is a collection of Decision Trees, most of the time trained with this bagging method. The general idea of this method is that a combination of learning models increases the overall results.

One of the advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems. But it cannot be applied as it's having problem in accuracy of prediction.

b)Artificial Neural Network(ANN):

An artificial neural network (ANN) is a computational model based on the derived structure and functions which is of biological neural networks. Informations that flows through the network affects the structure of the ANN. Because a neural network changes or learns, in a sense based on the input and output.

ANNs are considered nonlinear statistical data modeling tools where the complex relationships between the inputs and outputs are modeled or relationship patterns are found. ANN is also known as a neural networks. An ANN has several advantages but one of the most recognized of these is the fact that it can actually learn from observing data sets. In this way, ANN is used as a random function and approximation tool. These types of tools helps to estimate the most cost-effective and ideal methods for arriving at solutions while defining computing functions or distribution functions.

ANN is rarely used for predictive modelling. The reason being that Artificial Neural Networks (ANN) usually tries to over-fit the relationship. ANN is generally used in cases where what has happened in past or before is repeated almost exactly in same way.

c)Support vector machine(SVM):

SVM is one of the supervised ML classification techniques that is widely applied in the field of cancer diagnosis and prognosis. SVM functions by selecting critical samples from all classes known as support vectors and separating the classes by generating a linear function that divides them as broadly as possible using these support vectors.

Therefore, it can be said that a mapping between an input vector to a high dimensionality space is made using SVM that aims to find the most suitable hyperplane that divides the data set into classes [5]. This linear classifier aims to maximize the distance between the decision hyperplane and the nearest data point, which is called the marginal distance, by finding the best suited hyperplane [6].

SVM depends on the support vectors, which are the data sets closest to the decision boundary, in their algorithms. This is because removing other data points that are further away from the decision hyperplane will not change the boundary as much as if the support vectors were removed. The accuracy level and decision is based on the plotted points are the main advantages of this algorithm.

Here the comparison made on the three algorithms and the resultant graph is shown below:

The Figure 1 shows that the accuracy of SVM is much better than RF and ANN. However initially they fall from the same level and based on their prediction results SVM is applicable for further steps in prediction.

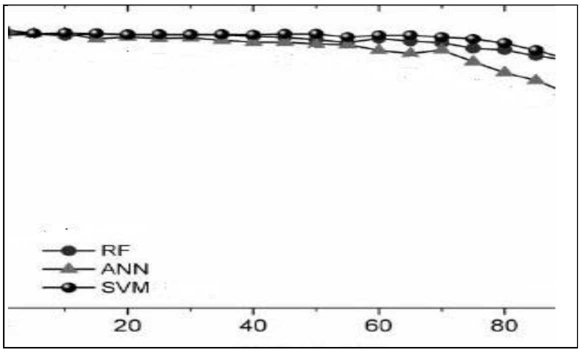

Figure 1. ***Comparison Result***

Hence from the above results the SVM is considered to produce output with accuracy.

3. INTELLIGENT MODEL FOR PREDICTION

This project have been implemented using R language in R tool. The module proposed for this project is listed below.

- Data Preprocessing
- Evaluation of SVM
- Result Prediction

The data preprocessing should be performed to take a valid dataset to the next step. There the algorithm which is selected for prediction is examined for its accuracy separately by producing confusion matrix. This matrix is used to conclude with two things

- Accuracy will be measured for selected algorithm.
- Prediction will be performed based on plotted points with respect to its plane.

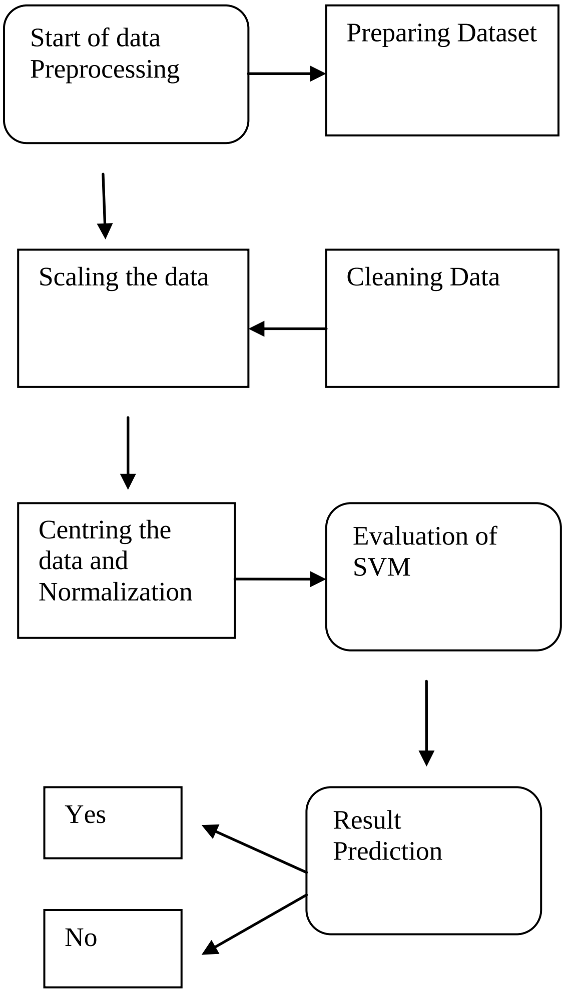

Figure 3. ***Flowchart of Breast Cancer Prediction***

a)Data Preprocessing

Data preprocessing involves data mining techniques in order to transform raw patient's data into required format, to be used by the classifier for detection and identification of the prediction patterns

(i)Business understanding: understanding the problem in breast cancer scope and thinking how to solve it, what we need

to solve this problem.

(ii)Collect data: understand breast cancer patient's data after collecting it from Gaza Strip hospitals and from MOH, and selecting target or relevant data based on the goal or data mining task.

(iii)Cleaning data: Data cleaning is a machine learning problem that needs data systems help. When dealing with real-world datas, dirty data is the normal rather than the exception. We continuously need to predict correct values, impute missing ones, and find links between the various data artefacts such as schemas and records. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of the data in analytics and decision making.

(iv)Scaling: Scaling is a machine learning method used to standardize the range of independent variables or features of data. In data preprocessing it is also known as data normalization and it is generally performed during the data preprocessing step**.** The main advantage of scaling is to avoid attributes which are greater in numeric ranges dominating those in smaller numeric ranges.

(v)Centring: Centring simply means subtracting a constant from every value of a variable. But when you Center X so that the value within the dataset becomes zero, the intercept becomes the mean of Y at the value you centered on it. Scaling is a method used to standardize the range of independent variables or features of datas. In data processing step, it is also known as a data normalization and is generally performed during the data preprocessing step.

Data collection is one of the most important phases in the research project. It included studying the underlying business, data understanding and gathering information from the team who are responsible for cancer disease in the Gaza Strip, sample of the dataset shown in Below.

The sample dataset will be,

- Clump Thickness

- Uniformity of Cell Size
- Uniformity of Cell Shape
- Marginal Adhesion
- Single Epithelial Cell
- Bare Nuclei
- Bland Chromatin
- Normal Nucleoli
- Mitoses
- Class

b) Evaluation of SVM

The support vector machine (SVM) approach is a classification technique based on Statistical Learning Theory (SLT). It is based on the idea of a hyperactive plane classifier, or linear separability.

(i)Generating Confusion Matrix:

A confusion matrix is a technique that is used for summarizing the performances of a classification algorithm.

Classification accuracy also can be misleading if you have an unequal number of observations in each class or if you have more than two classes in dataset.

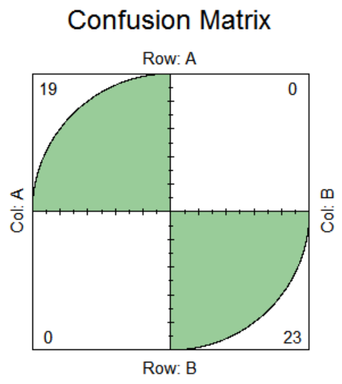

Figure 3. *Confusion Martix*

Calculating the confusion matrix can give a better idea of what classification is right and what type of error it is making.

The number of correct and incorrect predictions in the plotting are summarized with count values and broken down by each classes.

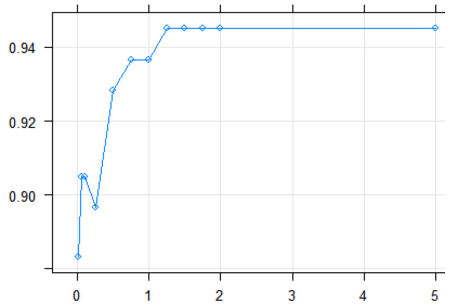

Figure 4. ***Accuracy Chart***

A Support Vector Machine (SVM) performs classification by finding the hyperplanes that maximizes or expands the margin between the two classes. The vectors (cases) that define the hyperplane are the support vectors [8] [7]. Suppose we have N training data points { (x1, y1),(x2, y2),(x3, y3), ...,(xN , yN) }, where xi belongs to Rd and yi belongs to { +1, −1 }. Consider a hyperactive plane defined by(w, b), where w is a weight vector and b is a bias. A new objectx can be classified with the following function:

$$f(x) = sign(w.x + b) = sign(\sum_{i=0}^{n} aiyi(xj, x) + b) \qquad \text{---(1)}$$

In practice, the data is often not linear capable of being linearly separated. However, one can still put into use a linear model by changing the data points via a non-linear mapping to another higher dimensional space (feature space) such that the data points will be capable of being linearly separated. The mapping is done by a kernel function and represented as k.

The nonlinear decision function of SVM is represented by the following function:

$$\text{f(x)=sign(}\sum_{i=0}^{n} aiyik(xi, x) + b\text{))} \qquad \text{---(2)}$$

Where k (xi, x) is the kernel function [8],
[9]

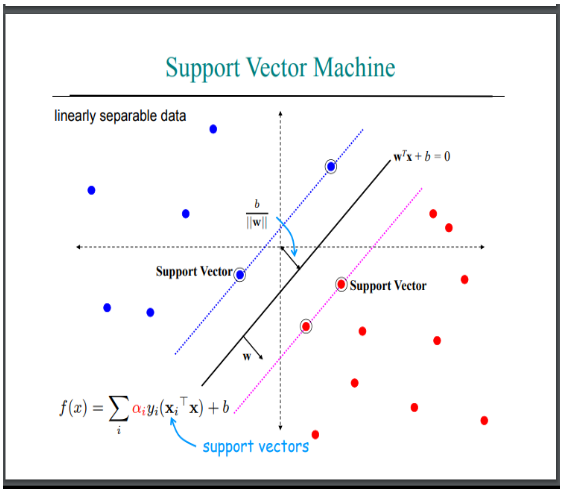

Figure 5. ***Linear SVM***

The linear SVM method is applied to classify the input points. Because it can efficiently handle the extra large data sets. There is no need for expensive computing data. The SVM linear method considers two classes (Class 0 and Class 1). The Class 0 shows that the particular person is healthy (not affected by cancer). Class 1 shows the people who are all affected by cancer. Because the thickness of cancer cells has a significant role in the degree of cancer severity. So if the thickness of cancer cells high then the degree of breast cancer severity high.

The dataset considered for processing will separate it as Training and Testing set. They are separated in the ratio of 70 and 30 respectively. After applying linear SVM based on main factors Nucleoli and

Mitoses in dataset the class will classify the planes and points. The result obtained after applying SVM is shown with two figures. Figure 6 is for Training set and Figure 7 is for Test Set. In order to separate two classes in plane different colours are applied in SVM. The number of points refers the number of records in dataset.

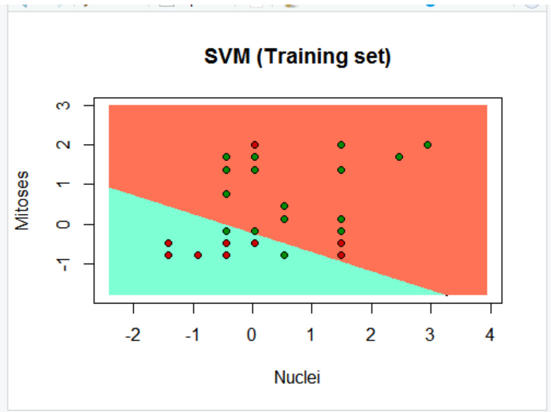

Figure 6. ***For Training Set***

In Training set number of people in larger area with green points refers to healthy person and other points in opposite plane refers the number of persons affected by cancer.

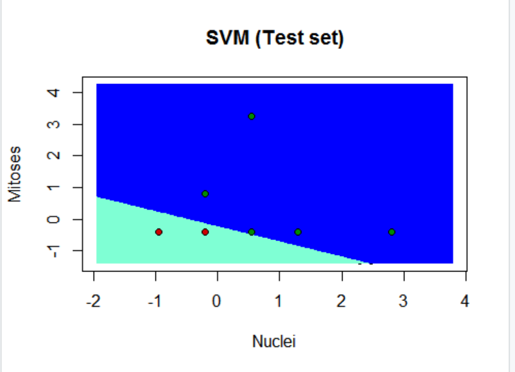

Figure 7. ***For Test Set***

Similarly this Test Set also explains result for test set which is selected from dataset.

4. CONCLUSION

In implementing the intelligent models, three major stages were involved in the development of intelligent severity prediction model, which include (I) data pre processing (II)Evaluation of SVM (III)Result Prediction. The data preprocessing used to eliminate the risks in dataset and simplify the relevant value in the dataset. The Evaluation of SVM calculates its accuracy and Result prediction produces the final result that how many people were affected and not affected by breast cancer using SVM Linear function. Additionally, here the comparison chart for accuracy of SVM, RF and ANN has been proposed. This work can be developed further by increasing the dimensionality of the dataset. Another enhancement is possible is new algorithm can be created or existing algorithm can be implemented to predict result more accurately. Based on level of severity the proper diagnosis will be suggested by the extension of this project.

REFERENCES

[1] Dr. Alaa M. El-Halees “Information Technology College Islamic University of Gaza Gaza, Palestine Breast Cancer Severity Degree Predication Using Data Mining Techniques in the Gaza Strip”

[2] Longo, D. L., Fauci, A. S., Kasper, D. L., Hauser, S. L., Jameson, J. L., & Loscalzo, J. (2012). “Harrison's Principles of Internal Medicine” 18E Vol 2 EB: McGraw Hill Professional.

[3] Duke, R. C., Ojcius, D. M., & Young, J. D.-E. (1996). “Cell suicide in health and disease”. Scientific American, 275(6), 80-87.

[4] Kim, S.-K., Yang, S., Seo, K. S., Ro, Y. M., Kim, J.-Y., & Seo, Y. S. (2005). “Home photo categorization based on photographic region templates”. Paper presented at the Asia Information Retrieval Symposium.

[5] G. Williams, “Descriptive and Predictive Analytics”, Data Min. with Ratt. R Art Excav. Data Knowl. Discov. Use R, pp. 193-203, 2011.

[6] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” Comput. Struct. Biotechnol. J., vol. 13, pp. 8-17, 2015.

[7] Heba, F. E., Darwish, A., Hassanien, A. E., & Abraham, A. (2010). “Principle components analysis and support vector machine based intrusion detection system”. Paper presented at the Intelligent Systems Design and Applications (ISDA).

[8] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition.”Data mining and knowledge discovery”, 2(2), 121-167.

[9] Kim, S., Shin, K.-s., & Park, K. (2005). “An application of support vector machines for customer churn analysis: Credit card case”. Advances in Natural Computation, 427-427.

[10] Bhavya, Mahak, & Mittal, P. (2015, 19-20 March 2015). “Data mining in medicine: Current issues and future trends. Paper presented at the Computer Engineering and Applications” (ICACEA).

[11] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). “From data mining to knowledge discovery in databases. AI magazine”, 17(3), 37.

[12] Olson, D. L., & Delen, D. (2008). “Advanced data mining techniques: Springer Science & Business Media”.

[13] Y. Elobaid, T.-C. Aw, J. N. W. Lim, S. Hamid, and M. Grivna, “Breast cancer presentation delays among Arab and national women in the UAE, a qualitative study,” *SSM - Popul. Heal.*, Mar. 2016.

[14] E. D. Michie, D. J. Spiegelhalter, and C. C. Taylor, “Machine Learning, Neural and Statistical Classification,” *Proceeding*, 1994.

[15] T. Fushiki, “Estimation of prediction error by using K-fold crossvalidation,” *Stat. Comput.*, 2011.

[16] K. J. Edwards and M. M. Gaber, “Astronomy and Big Data”.2014

[17] D. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation,” J. Mach. Learn. Technol,2011.

[18] L. F. Carvalho, G. Fernandes, M. V. O. De Assis, J. J. P. C. Rodrigues, and M. Lemes Proenc.a, “Digital signature of network segment for healthcare environments support,” Irbm, 2014.