International Journal of Research in Science and Technology http://www.ijrst.com

(IJRST) 2017, Vol. No. 7, Issue No. III, Jul-Sep **e-ISSN: 2249-0604, p-ISSN: 2454-180X**

CANCER DETECTION USING MACHINE LEARNING: A GENERALIZED APPROACH

Ayush Sharma

Department of Computer Science and Engineering
Jaypee Institute of Information Technology
Noida, India

ABSTRACT

Accurate prediction of cancer can play a crucial role in its treatment. The procedure of cancer detection is incumbent upon the doctor, which at times can be subjected to human error and therefore leading to erroneous decisions. Using machine learning techniques for the same can prove to be beneficial. Many classification algorithms such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN) are proven to produce good classification accuracies. The following study models data sets for breast, liver, ovarian and prostate cancer using the aforementioned algorithms and compares them. The study covers data from condition of organs, which is called standard data and from gene expression data as well. This research has shown that SVM classifier can obtain better performance for classification in comparison to the ANN classifier.

Keywords—Support Vector Machine, Artificial Neural Network, Cancer, accuracy, machine learning.

INTRODUCTION

Cancer is the name given to a collection of related diseases. The general infiltration process of cancer is similar irrespective of the type of cancer: cancer cells cause body cells to divide uncontrollably which then spreads to surrounding tissue. Cancer inception can consequently begin anywhere in the human body, which is comprised of uncountable cells that make it up. The usual cell cycle is as follows: all human cells, in fact every living cell grows and undergoes division to form new cells by the process of mitosis. When cells grow older, or their existence is hampered by some foreign body, they become dead and new cells supplant them [3]. The sole reason behind early diagnosis and prognosis of cancer is that it facilitates further clinical procedures involved and helps improve the same for future patients. Furthermore, this helps classify patients into two broad classifications of high and low risk groups, which different procedures and methods for different groups. As a result, the aforementioned techniques help in providing a panacea for all cancer diseases [3]. The central aim of this thesis is to identify the unique genes – called as Driver Genes – pertinent to causing cancer, of any subtype. The more driver genes interpreted, the more the horizons to explore of the medication of treating cancer. But identifying a specific driver is an arduous task, mostly in the

cases where the complete chromosome is perished. Orthodox and conventional biological methods are subjected to torpor and inaccurate, having identified only a small amount. As cancer "big data" concocts, researchers are digressing to computational methods, specifically machine learning, to identify additional drivers. IT is still confounding how to use these methods for the application of biomedical sciences and prediction of carcinoma. Cancer itself can be further divided into: Breast, Prostate, Ovarian and Lung, liver etc. Even after such granulation each of its subtype is completely anomalous and unique. Because of the very reason mentioned above, it difficult to model all possible cancer types. The crux of all cancers is tantamount in fact: a gene that goes haywire, inside a chromosome and starts disrupting bodily functions by dividing ceaselessly.

Cancer cells are the ones that are subjected to rapid division, by mitosis, and evolution, inside the very human body, which makes identifying the driver gene all the more difficult. The degree of aberration with cancer cells is to such an extent that their entire DNA integrity and error correction is out of order, resulting in the accumulation of multiple specious mutations. There are times when the very driver genes become perished, and leads to difficulty in ousting the passenger genes, which do not contribute in cancer [4].

Various supervised classification algorithms are available and the ones used here are Support Vector Machine and Artificial Neural Network. Both of the above algorithms are proven to produce good classification accuracy performance, however the results are incongruous as being highly dependent on the data sets. Due to the incongruous result obtained, the cardinal aim of this study is to further validate the performance of both Artificial Neural Networks and Support Vector Machines in cancer classification. The performance of both classifiers is tested and evaluated on four different cancer datasets which are divided into two type of cancer data namely; standard data, which is obtained from condition of various organs in the body and second being the gene expression data. These four datasets are obtained from UCI Machine Learning Repository [4] and National Cancer Institute (NCI) [5]. The paper is categorized in the following manner: subsequent section delineates the dataset in depth, divided namely into standardized and gene expression data. It also explains how different features affect the model as a whole in terms of performance and accuracy. Section 3 explains the algorithms used in this study. Finally, section 4 enlists the results and observations of this study and finally ending the study with the future scope of machine learning applications in cancer detection in section 5.

LITERATURE REVIEW

A. Related Work

The research using machine learning techniques in the medical domain has been prevalent for a long time, especially in the diagnosis of cancer. Various researches have been conducted, using an approach similar to the one delineated here. A similar research was carried out by author Ismail Sartias in [6] where they were able to diagnose cancer severity using Artificial Neural Network on a data set of 800 patients obtained from Biopsy results from research conducted by M. Elter, R. Schulz-Wendtland and T. Wittenberg. The two conclusive results obtained, namely the diseases prediction

ratio of 90.6% and a health ratio of 80.9% could ultimately be conducive to oncologists and physicians, establishing artificial neural networks as a robust model with good reliability.

A similar approach was proposed by Alvarez and Sanchez in [7] where they use various materials and methods for the task of classification, preprocessing and evaluation. For data preprocessing, it uses dimensionality reduction. Then, it uses techniques like SOM (Self-organizing maps), Support Vector Machines, MARS (Multivariate adaptive regression splines) and Neural networks for modelling. Finally, for the task of evaluation, it uses the metric sensitivity and specificity to asses which model performs the best. Specificity of 94.5% from MARS model was obtained whereas an accuracy of 86.1% in SVM classifier.

Numerous studies have been engendered by various researchers in order to classify cancer using sophisticated classifiers. However, the result obtained from prior researches are inconsistent. Some studies state that ANN is better compared to SVM. One such example is a research conducted by [8] using models like Support vector machines and probabilistic neural networks (PNN), a type of ANN and achieving a definitive result with high accuracies from PNN. [9] also found that ANN outperforms SVM in the classification of Micro-Calcification Clusters (MCCs) in mammogram imaging. Another research, elucidated in [10] juxtaposes polynomial SVM with Radial Basis Function Neural Network and states that the latter performed better to classify breast tumors, on the Wisconsin breast data set. Due to this discrepancy, the research aims to substantiate the result further.

B. Artificial Neural Neural Network

Artificial Neural network is the subfield of machine learning and computational intelligence that aims to learn from patterns, data entries and attributes from historical data and learns according to them and improves from every occurrence to make prediction. Human beings are comprised of neurons in brain and ANN's are analogous to them in some sense. Similar to the human brain, it comprises of highly structural and convoluted interconnected network of entities called nodes/units, where each unit mimics the biological equivalent of neuron. The nodes/units are provided with a weighted set of inputs and give outputs based on them [13]. Figure 10 reveals how different nodes work in Artificial Neural Networks. The most widely used and acknowledged model of ANN is the Back Propagation Neural Network (BPNN), also known as multi-layer feed-forward neural networks [9,14]. The abovementioned type of ANN is based on supervised learning algorithm, which means that it can only learn from data which is labeled and give output for the same [14]. The algorithm for BPNN is simple; signals travel from input neuron to output neuron without returning to the source itself, therefore reducing overhead [9]. The BPPN structure at least consists of three layers; which are the input layer, the output layer and the intermediate hidden layer. The number of nodes are different for different layers, depending on the number of variables. For example, for the input layer the number of nodes will be tantamount to the number of input variables, where in case of binary classification, number of nodes in output layer would be two only (0 or 1). For cancer classification, it would be either benign: 0, or malignant: 1.

The BPNN is called so because of a simple phenomenon; the error is first calculated at the output layer, from where it is back propagated into the hidden layer and finally the input layer. The above

phenomenon is completed in two steps at most: the forward activation that produces a solution, and the backward propagation that produces the error to alter the weights. These steps are carried out ceaselessly until the ANN model agrees with the desired value within the model tolerance. [9].

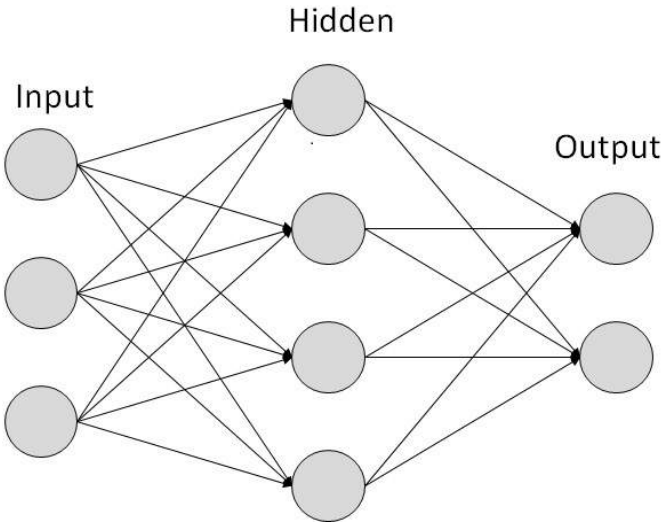

Fig. 1 A simple neural network model representation

C. Support Vector Machine Classifier

The support vector machine classifier, which is predicated on statistical learning theory, is a technique for supervised learning, i.e. when the data is labeled. The model was conceived firstly by Cortes and Vapnik in their initial research conducted in [13]. The algorithm is straight forward; it fits a hyperplane on the given data and checks how accurate the hyperplane fits with a good classification accuracy on either side of the hyperplane. The aforementioned case is when data is not linearly separable, i.e. it is required to segregate the data into two or hyperplanes for distinct classification, whereas it fits a straight line when the data itself is linearly separable. This is elucidated much more articulately in fig. 8 and 9. The support vector machine revolves around creating a margin for the fitted data; maximizing the margin leads to greater accuracy and vice versa reduces error by a fair share. The linear SVM minimizes the loss function:

$$f_i(\theta) = \max (1 - \theta^T x, 0)$$

SVM behaves as a linear classifier in the feature space of linearly separable data. On the contrary, it manifests into a nonlinear classifier by instead fitting a straight line rather than fitting a hyperplane, as a result of linearly separable data from input space.

EXPERIMENTAL DATA

As mentioned above, the study here encompasses four different cancer data sets, namely breast, liver, prostate and ovarian cancer data sets. The reason behind choosing four disparate data sets is to eliminate the discrepancy where a model tends to favor a particular kind of data set. Consequently, the data sets can be segregated into two distinct classes, one where the features of the data are

obtained from organ conditions, also called standard data, and the other one being gene expression data. Breast and Liver cancer fall under the category of standard data, whereas prostate and ovarian cancer fall under the gene expression data class. The breast cancer data is obtained from UCI Machine Learning repository [4], and is named as Wisconsin Diagnostic Breast Cancer dataset [15] [16]. The BUPA Medical Research Limited provides with the data for liver cancer which is elucidated below. This dataset was firstly conceived by Richard S. Forsyth in 1990. Both the breast cancer and liver data sets are described in the Table 1 and 2 respectively.

The diagnostic data set has 458 entries with benign tumor and 241 with malignant tumor. It has the following features:

TABLE 1 Wisconsin Diagnostic Breast Cancer data set description

S no.	***Attribute/Feature***	***Range***
1.	ID Number	Identification number for patients
2.	Diagnosis	2: Benign, 4: Malignant
3.	Radius	11-27
4.	Area	360-2300
5.	Perimeter	71-82
6.	Texture	11-40
7.	Smoothness	0.05-0.2
8.	Compactness	0.04-0.45
9.	Concavity	0.02-0.5
10.	Concave Points	0.02-0.5
11.	Symmetry	0.1-0.3
12.	Fractal Dimension	0.05-0.1

The liver disorders data set, produced by BUPA Medical Research Ltd comprises of a total 7 attributes, with 345 total number of instances and no missing values. There are some attributes which are highly sensitive to liver disorders pertaining to increased alcohol consumption. The first 5 attributes encompass those attributes. Information about the same is provided below:

TABLE 2 BUPA Liver Disorders data set

S no.	***Attribute/Feature***	***Full Form***
1.	MCV	Mean Corpuscular

S no.	***Attribute/Feature***	***Full Form***
		Volume
2.	Alkphos	Alkaline Phosphatase
3.	SGPT	Alamine aminotransferase
4.	SGOT	Aspartate aminotransferase
5.	gammagt	Gamma-Glutamyl Transpeptidase
6.	drinks	Number of alcoholic beverages drunk per day that are equivalent to half-pints.
7.	Selector	Attribute to bifurcate the dataset

The prostate cancer dataset that goes by the name of JNCI Data in the repository of National Cancer Institute, dated 7-03-2002, comprises of serum spectra results of 322 with a peak amplitude reaching a 1514 points following in the range of 0 – 2000 Da. The range defines the ration of mass by protein present in the chromosome at a particular time. There are 253 benign and 69 malignant samples in the dataset. The ovarian cancer dataset is labelled as "Ovarian 8-7- 02", and consists of 253 datasets. The spectra comprise of 91 benign samples and 162 malignant samples. Gene expression data is one which is defined as the flow of genetic information from gene to protein, and includes the prostate and ovarian cancer datasets in the category. A data or commonly called as a mass spectrum in the context of gene expression data contain thousands of different mass/charge ratios. For both of the dataset, each data contains 15154 values of m/z in the range of 0-20000 Da. These values are then called as features/labels in this study.

TABLE 3 Summary of data sets used

S no.	***Data set***	***Instances***	***No. of features***
1.	WBCD	699	9
2.	BUPA Liver disorder	345	6
3.	JNCI 7-3-02	322	15154
4.	Ovarian 8-7-02	253	15154

METHODOLOGY

Diagnosis of cancer is an arduous task. The first step is the clinically examination to detect the tumor/lump in the affected organ by the General Surgeon or by using imaging techniques such as Mammography, Pap Test, Biopsy etc. The results of these imaging test provide the features for that respective cancer which can be further used for modelling using machine learning techniques. Finally,

models such as SVM and ANN are trained using the obtained data to make predictions. The general methodology of modelling after data is obtained can be divided into four general steps which can be shown in the figure below. No matter what machine learning algorithm is being implemented, the aforementioned steps are key to any algorithm.

D. Data Preprocessing

This is the first step before modelling data. At times data obtained may be incomplete, like lacking values in certain rows. It can also be noisy, replete with errors and outliers, and inconsistent as well, with arrant discrepancies. Thus, in order to eliminate all of the above, we preprocess the data before modelling. Here, we use min max normalization, also known as feature scaling, since the entries inside our data are primarily numeric. It can be delineated using the formula below [18]:

$$y = \frac{x - min(x)}{max(x) - min(x)}$$

where x defines the initial value and min and max are the biggest and lowest values respectively.

E. Data Visualization

Data Visualization is the disposition of data pictorially or graphically using a thirdparty application. The prime of importance of data visualization lies in identification of patterns or trends in data, how to weigh individual features, and how to identify outliers in data. Furthermore, it ultimately helps in selection of right model for modelling of data. Before visualizing data, it is important to understand the size as well as the cardinality of data. High cardinality means there's a larger percentage of unique values, whereas converse is true for data with low cardinality. Secondly, it is important to determine what you're trying to visualize.

There are numerous ways to visualize data. The most common ones are bar graphs, line charts, box plots, and heat maps etc. The same will be discussed further in the paper.

F. Model Selection and Implementation

The choice of selecting the right model is entirely subjective. It depends strongly on the type of data, and what is the primary aim of the author. If primary aim is accuracy, the best bet is to test data on number of models and then select the best one using cross validation. However, when the need is of a good enough model, there are certain things to be kept in mind. Firstly, the size of training set plays an integral role. The high bias or low variance classifiers have an advantage over their counterparts, that is low bias or high variance classifiers in case when the data entries in the training set are less, as the former is less prone to overfitting data. Each model has its own disadvantages and advantages. For example, using Logistic Regression offers many ways to regularize the model, and eliminates the constraint of features being correlated. Although SVM's are known to work well on linearly separable data, with a suitable kernel they can work on non-linear feature space as well. The data set is divided further into two sets: training and test set. The split is usually done using the ratio of 80/20, which means 80% of data is used for modelling and 20% of data is used for evaluation and prediction. Thus, we select here four models as discussed earlier, and divide both data sets into training and testing sets.

G. Model Evaluation

Once the training data is modelled, we evaluate the test data and predict the outcome. The labels of test data are recorded and the incorrectly predicted labels are counted, giving us the simplest form of evaluation of model. The cardinal aim of any machine learning algorithm, predominantly classification algorithms, is to classify unseen data and predict it into the correct class. We assume that our samples are independent and identically distributed, which means that all samples have been drawn from the same probability distribution and are statistically independent from each other.

The performance of the classifiers was evaluated using metrics such as accuracy, sensitivity, specificity, and area under curve (AUC) etc. Sensitivity of a model is defined as how many true entries were predicted actually as true by the model, in this case the entries being the tumor. The classifier that can correctly classify benign tumors will have a higher result in sensitivity. Sensitivity is defined as follows [19, 20]:

$$\text{Sensitivity (\%)} = \text{TPR} = \frac{\text{TP}}{\text{FN} + \text{TP}} \text{X100}$$

Specificity is the percentage of malignant tumors data classified as malignant by the classifiers. The classifier that can correctly classify malignant tumors will have a better result in specificity. Specificity is generally the number of false entries actually predicted as false by the model, in this case the false entries being the malignant tumors which is calculated as follows [21, 22]:

$$\text{Specificity (\%)} = \text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \text{X100}$$

Accuracy combines the terminology of specificity and sensitivity and then compares the results with respected to the total number of data entries. Higher value for accuracy is indicative of a better performance of the model. It is given by [23]:

$$\text{Accuracy (\%)} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{FP} + \text{TP} + \text{FN}} \text{X100}$$

Area Under Curve (AUC) establishes particular thresholds while evaluating specificity and sensitivity. The AUC value of 100% represents perfect discrimination (the classifier can classify the tumors correctly), whereas an AUC value of 50% is equivalent to random model. AUC was calculated as follows [20 here, 15]:

$$\text{AUC (\%)} = \frac{1}{2}\left(\frac{\text{TN}}{\text{TN} + \text{FP}} + \frac{\text{TP}}{\text{TP} + \text{FN}}\right) \text{X 100}$$

Although there are other metrics for model evaluation available as well, classification accuracy is the one most pertinent to our research.

RESULTS AND DISCUSSION

Throughout the research, two models were closely scrutinized and evaluated for four different datasets using the aforementioned metrics such as sensitivity, specificity, accuracy, and area under

curve. The result from confusion matrix for each model is formulated in order to evaluate the models with ease.

For the Wisconsin breast cancer diagnostic data set, the results were shown in Table 4 below. Similarly, the results after modeling the liver cancer data set are shown in table 5. Furthermore, there appears a common trend whilst modeling the standard data sets. The support vector machine tends to perform better for standard set than for gene expression data in classification. Similarly, the results for classifiers on JNC 7-3-02 and Ovarian cancer datasets are shown in table 6 and table 7 respectively.

TABLE 4 Classifier results on WDBC

Metric.	***Support Vector Machines***	***Artificial Neural Networks***
Specificity	98.2%	94.7%
Sensitivity	93.22%	89.08%
Accuracy	96.66%	93.06%
AUC	99.63%	92.39%

TABLE 5 Classifier results on BUPA

Metric.	***Support Vector Machines***	***Artificial Neural Network***
Specificity	99.86%	32.56%
Sensitivity	36.67%	75.10%
Accuracy	63.12%	57.31%
AUC	68.34%	52.76%

The reason why SVM performs better for standard data than for gene expression data is because ANN tends to converge on local minima rather than global minima and often over fits if training goes long, which translates into noise being considered as a part of the pattern itself.

Conversely, we see that ANN tends to perform better than SVM for gene expression data. This may be plausible for couple of reasons: ANN is a parametric model, whereas SVM is non-parametric model. In worst case, the number of support vectors in SVM may be equal to number of training examples, and therefore the model size scales linearly. On graphically comparing performance of each of the classifier based on accuracy shown in Fig. 2, it is evident that due to imbalanced spread of tumors for ovarian and prostate cancer, the accuracy struggles in case of SVM for gene expression data, whereas less number of features in standard data set affect the accuracy for ANN.

TABLE 6 Classifier results for JNC 7-3-02

Metric.	***Support Vector Machine***	***Artificial Neural Network***
Specificity	0.00%	19.01%
Sensitivity	100.00%	100.00%
Accuracy	78.34%	82.37%
AUC	50.02%	59.31%

TABLE 7 Classifier results on Ovarian

Metric.	***Support Vector Machine***	***Artificial Neural Network***
Specificity	100.00%	100.00%
Sensitivity	0.00%	40.73%
Accuracy	64.47%	78.97%
AUC	50%	70.37%

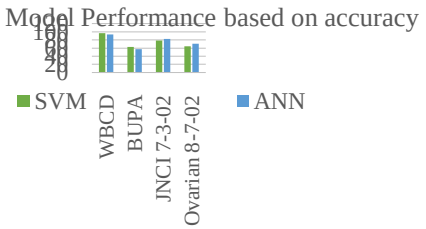

Fig. 2 Model performance based on accuracy only

CONCLUSION

Accurate prediction of cancer still remains an arcane task. While The scope of Machine Learning in biomedical sciences is not confined to using only sophisticated classification algorithms for binary classification. Cancer is a perilous disease and it is not limited to a two-step diagnosis or prognosis procedure using physical examination or biopsy. For any type of cancer, after physically examining the lump/mass, the next step is to further validate it using imaging techniques. Instead of first diagnosing the tumor from the imaging result and then finding the test result on that respective tumor can be eliminated by implementing the machine learning techniques on the images itself. Here,

unsupervised learning comes to rescue, where we can use various methods such as clustering, Neural Networks, Deep learning etc. Assigning weights to tumor points on the imaging result itself and then calculating the weighted score for each image result helps in establishing a firm ground for cancer diagnosis. Here, we can eliminate the ternary procedure of FNA and using features from FNA to model classifiers.

Eventually, we can extend these techniques to hospitals, and maintain a repository of patients with their current diagnosis, and every other detail, which helps in accounting for new cases and immediately checking the database for similar patients using the same Machine Learning techniques. The application of Machine learning in wide gamut of fields has been inexorable and if learned how to harness it to its true potential, machine learning can give groundbreaking results.

REFERENCES

[1] Nation Cancer Institute (NCI), *Cancer – A detailed Guide* http://cancer.gov/about-cancer, Accesed on March, 2016.

[2] Konstantina Kourou, Themis P. Exarchos, Michalis V. Karamouzis, Machine learning applications in cancer prognosis and prediction, accesed on August, 2016.

[3] Ahrim Youn, Richard Simon, Identifying cancer driver genes in tumor genome sequencing studies, Bioinformatics. 2011 Jan 15; 27(2): 175–181.

[4] UCI Machine Learning Repository of data sets for machine learning, http://archive.ics.uci.edu/ml/, Accesed at July, 2016

[5] Data Catlog, Nation Cancer Institute R&D Resources, https://www.cancer.gov/research/resources/data-catalog, Accesed July 2016.

[6] Ismail Saritas, Prediction of Breast Cancer using Artificial Neural Networks, Springer Science+Business Media, LLC 2011, Published on 12 August, 2011.

[7] AS Ren J. 2012. ANN vs. SVM: Which One Performs Better in Classification of MCCs in Mammogram Imaging. *Knowledge-Based Systems*. 26: 144–153. 10 (1), January 2012.

[8] I.S. Subashini T. S., V. Ramalingam, and S. Palanivel. 2009. Breast Mass Classification Based on Cytological Patterns using RBFNN and SVM. *Expert Systems with Applications*. 36: 5284–5290.

[9] Ojha, Varun Kumar; Abraham, Ajith; Snášel, Václav (2017-04-01). "Metaheuristic design of feedforward neural networks: A review of two decades of research".

[10] Bottaci, Leonardo. "Artificial Neural Networks Applied to Outcome Prediction for Colorectal Cancer Patients in Separate Institutions".

[11] Cortes, C.; Vapnik, V. (1995). "Support-vector networks". *Machine Learning*. 20 (3): 273–297.

[12] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

[13] William H. Wolberg and O.L. Mangasarian*: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.*

[14] A Pär Stattin, Sigrid Carlsson, Benny Holmström, Andrew Vickers, Jonas Hugosson, Hans Lilja, Håkan Jonsson, Pro Prostate Cancer Mortality in Areas With High and Low Prostate Cancer Incidence https://academic.oup.com/jnci/article/doi/10.1093/jnci/dju007/1745564/Prostate-Cancer-Mortality-in-Areas-With-High-and.

[15] Liu Y., and Y. F. Zheng. 2004. FS_SFS: A Novel Feature Selection Method for Support Vector Machines. IEEE International Conference on Acoustic, Speech, and Signal Processing. 5: 797–800.

[16] Keyvanfard F., M. A. Shoorehdeli, and M. Teshnehlab. 2011. Feature Selection and Classification of Breast Cancer on Dynamic Magnetic Resonance Imaging using ANN and SVM. American Journal of Biomedical Engineering. 1: 20–25.

[17] Subashini T. S., V. Ramalingam, and S. Palanivel. 2009. Breast Mass Classification Based on Cytological Patterns using RBFNN and SVM . Expert Systems with Applications. 36: 5284–5290

[18] Ren J. 2012. ANN vs. SVM: Which One Performs Better in Classification of MCCs in Mammogram Imaging. Knowledge-Based Systems. 26: 144–153.

[19] Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". *The American Statistician*. **46** (3): 175–185.