International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 06, June 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

# Survey on Breast Cancer Analysis using Machine Learning Techniques

**Prof Tejal Upadhyay[1], Arpita Shah[2]**

[1]*Assistant Professor, Information Technology Department,*

[2]*M.Tech, Computer Science and Engineering, Specialization in Network Technology*

*Department of Information Technology, Institute of Technology,*

*Nirma University, Ahmedabad, Gujarat (India)*

## ABSTRACT

*In this paper we have use different data mining techniques for Prediction of Breast Cancer Survivability. The use of machine learning and methoding techniques has revolutionized the full process of carcinoma prognosis. In this paper we present a survey of those models that are being used to enhance the breast cancer prognosis prediction. We have introduced these models and secondly we have given an overview of the current research being carried out using these models. We specify different level of accuracies being claimed by different researchers. Lastly, we conclude that despite the ongoing research efforts towards achieving better capabilities for prediction system, we still need much more to build a more accurate and less invasive prognostic system that can benefit the mankind. This paper consists of survey of published papers with various techniques comparison like artificial neural networks, Support Vector Machines, Fuzzy logic and many more. It consists of various parameters and metrics on which the comparisons are made so that there can be an idea to use a technique for predicting Breast Cancer Survivability. Data analysis systems, intended to assist a physician, are highly desirable to be accurate, human interpretable and balanced, with a degree of confidence associated with final decision. In cancer prognosis, such systems estimate recurrence of disease and predict survival of patient.*

## I. INTRODUCTION

### 1.1 What is functional Genomics

Functional genomics is the study of how genes and intergenic regions of the genome contribute to different biological processes. The goal of functional genomics is to determine how the individual components of a biological system work together to produce a particular phenotype. Functional genomics focuses on the dynamic expression of gene products in a specific context, for example, at a specific developmental stage or during a disease A basic side of machine genomics is that the drawback of purposeful identification. All genes in associate organism have a particular operate or functions. This operate describes the role that a factor plays within the cell and is mostly determined by experimentation. Given a coaching set of genes with notable options and performance, a model may be made victimization machine learning to predict the operate of all the genes within the organism.[1] There square measure many specific genomics approaches looking on what we have a tendency to square measure targeted on:

I. DNA level (genomics and epigenomes)
II. RNA level (transcriptomics)
III. protein level (proteomics)

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

IV. metabolite level (metabolomics)

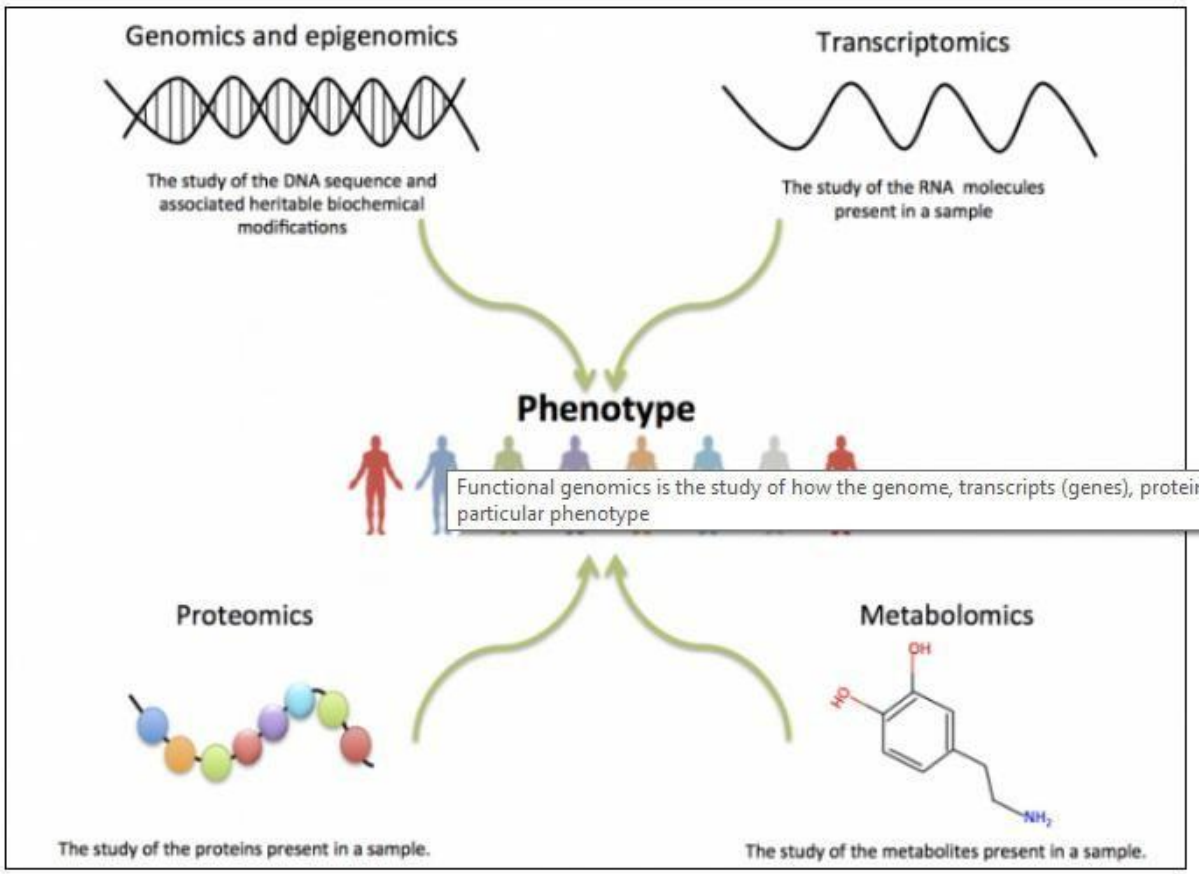


**Fig 1, Functional Genomics 1**

Examples of biological questions that can be tackled using functional genomics experiments are

1. Why do some cancer drugs only work effectively on a subset of patients with the disease?
2. Why are some cultivars of rice more resistant to drought than others?
3. What makes some individuals more susceptible to skin allergies?

Functional genomics experiments measure changes in the DNA (genome and epigenome), RNA , or interactions between DNA/RNA and proteins that influence the phenotype of a sample. Common branches of functional genomics include:

- Transcription profiling
- Genotyping
- Epigenetic profiling
- Nucleic acid-protein interactions
- Meta-analysis

**1.1.1 Transcription profiling:**

It is also known as 'expression profiling. It involves the quantification of gene expression of many genes in cells or tissue samples at the transcription (RNA) level. The quantification can be done by collecting biological samples and extracting RNA (in most cases, total RNA) following a treatment or at fixed time-points in a time-series, thereby creating 'snap-shots' of expression patterns[2]. In the field of biology, organic phenomenon

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

identification is that the measuring of the activity (the expression) of thousands of genes quickly, to make a world image of cellular perform.

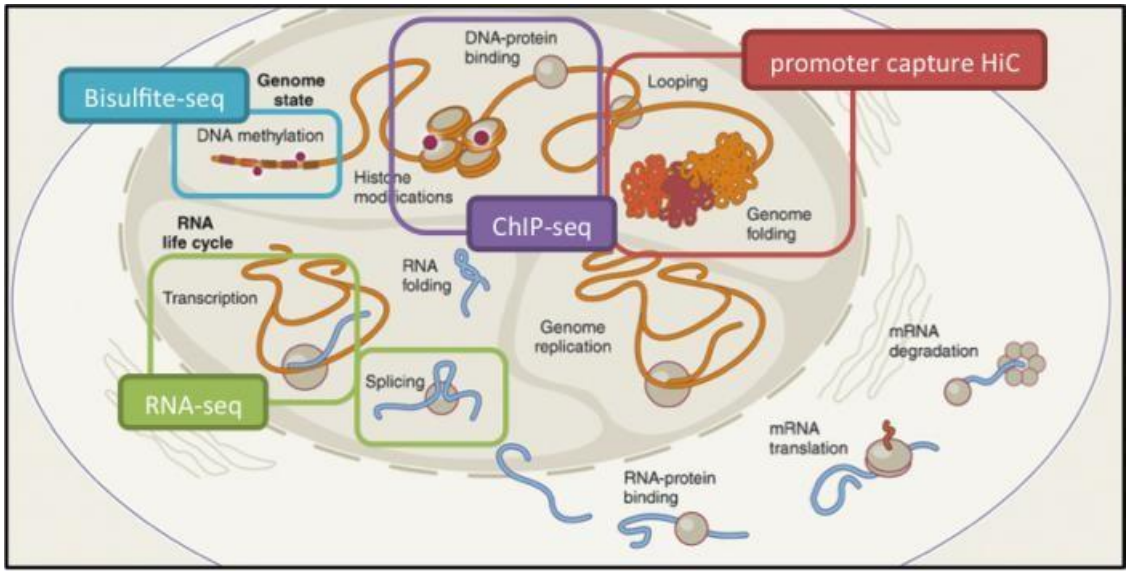


**Fig 2 Expression Pattern**

**1.1.2 Genotyping**

Genotyping is that the method of deciding variations within the genetic make-up (genotype) of a personal by examining the individual's DNA sequence exploitation biological assays and comparison it to a different individual's sequence or a reference sequence. It reveals the alleles a personal has familial from their folks. Genotyping studies are often designed to spot DNA sequence variations at 3 levels:

**Single nucleotide polymorphisms (SNPs, pronounced snips):** SNP analysis focuses on differences in the DNA sequence at the single nucleotide level.

**Copy number variations (CNVs**): CNVs refer to an increase or decrease in the number of copies of a segment of DNA.

**Structural variations:** they are an order of magnitude larger than CNVs and often cover mega bases of DNA, and can be caused by chromosomal rearrangement events.

**1.1.3 Epigenetic profiling :**

Epigenetics is the study of how biochemical modifications or physical interaction of DNA/chromatin affect gene regulation in a cell, where such modifications / interactions are not related to changes in the underlying DNA sequence. At the DNA level, methylation of CpG dinucleotides (often located near gene promoters) can be detected by first converting un methylated cytosine into uracil using bisulfite, which allows methylated and un methylated cytosine to be distinguished. At the chromatin level, modifications of the tails of histone proteins (e.g. methylation, acetylation) can be mapped by 'immuno-precipitation', where chromatin and proteins are chemically cross-linked reversibly. The genomic DNA associated with the modification/protein of interest is then 'pulled-down' with specific antibodies raised against the modification/protein.

**1.1.4 DNA/RNA-protein interactions:**

Transcription factors, ribosomes and other DNA/RNA-binding proteins can bind to nucleic acid sequences and influence the transcription and translation of genes. The immunoprecipitation technique has also been applied to study protein binding sites on RNA.

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

1.1.5 Meta-analysis:

Meta-analysis is a branch of functional genomics in which data from pre-existing experiments is combined to create statistically more powerful models of a biological process. This type of analyses has become popular as it allows the identification of subtle events that could not be detected in smaller studies. Functional genomics databases such as Array Express and Expression Atlas play an important role in these studies as reliable, well annotated sources of functional genomics data.

### 1.2 Breast Cancer Analysis

Breast cancer is that the second commonest reason behind deaths from cancer among girls within the us. In 2006, it's calculable that regarding 212 000 new cases of invasive carcinoma are going to be diagnosed, along side fifty eight 000 new cases of non-invasive carcinoma and forty 000 girls area unit expected to die from this illness (Data from yank Cancer Society, 2006). the foremost clinical downside of carcinoma is that the repeat of therapeutically resistant disseminated illness. distinctive a factor signature victimization microarray knowledge for carcinoma prognosis has been a central goal in some recent large-scale preliminary studies. In vant Veer et al., 2002, a 70-gene signature (also referred to as the capital of The Netherlands signature) was derived from a cohort of seventy eight carcinoma patients, the prognostic worth of that was additional valid during a larger dataset (van Delaware Vijver et al., 2002). additional recently, a 76-gene signature was known and with success wont to predict distant metastases of humour node-negative primary carcinoma (Wang et al., 2005).In Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Health care, According to National Cancer Institute of United States, estimated number of breast cancer cases, registered for the year 2007 is 180510, while the estimation of deaths exceeds 41000. Approximately, at the rate of one in three cancers diagnosed, breast cancer is the most frequently diagnosed cancer in women in America Prognosis helps in establishing a treatment plan by predicting the outcome of a disease.

There are three predictive foci of cancer prognosis: 1) prediction of cancer susceptibility (risk assessment), 2) prediction of cancer recurrence and 3) prediction of cancer survivability. Focus of this paper is prediction of survivability, of a particular patient suffering from breast cancer, over a particular time period after the diagnosis. In Predicting carcinoma survivability: a comparison of 3 data processing strategies , we have a tendency to used 2 well-liked data processing algorithms (artificial neural networks and call trees) in conjunction with a most ordinarily used statistical procedure (logistic regression) to develop the prediction models employing a massive dataset (more than two hundred,000 cases). we have a tendency to conjointly used 10-fold cross-validation strategies to live the unbiased estimate of the 3 prediction models for performance comparison functions. In this paper, we tend to report on our research wherever we tend to developed models that predict the survivability of diagnosed cases for carcinoma. one in all the salient options of this endeavor is that the believability and also the giant volume of knowledge processed in developing these survivability prediction models.

In Prognostic/Predictive Factors in carcinoma Shahla Masood, MD we tend to study Prognostic/predictive factors, node standing, Tumor type, DNA ploidy and proliferation rate. In wFDT - Weighted Fuzzy call Trees for Prognosis of carcinoma Survivability, Cancer prognosis estimates return of malady and predict survival of patient; therefore leading to improved patient management.

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

To develop such data primarily based prognostic system, this paper examines potential sexual union of accuracy and interpretability within the type of mathematical logic and call Trees, severally. result of rule weights on fuzzy call trees is investigated to be another to membership operate modifications for performance improvement.

Experiments were performed victimization totally different mixtures of: variety of call tree rules, styles of fuzzy membership functions and illation techniques for carcinoma survival analysis. SEER carcinoma knowledge set (1973-2003), the foremost comprehensible supply of data on cancer incidence in u. s., is taken into account. Performance comparisons recommend that predictions of weighted fuzzy call trees (wFDT) ar additional correct and balanced, than severally applied crisp call tree classifiers; what is more it's a possible to adapt for vital performance sweetening.

## II. METHODS

Prediction Models for Breast Cancer Survivability

### 2.1 Prediction Models :

Majorly following algorithms are used in prediction models for breast cancer survivability:

1) Decision Trees

2) Artificial Neural Network

3) Logistic Regression

4) Genetic Algorithms

5) SVM and some models are hybridization of these algorithms.

#### 2.1.1 Decision Trees

In Medical research domain decision trees are one of the powerful classification algorithms. A decision tree is a tree like graph suggesting decisions and their possible consequences. Popular decision tree algorithms include Quinlans ID3, C4.5 and C5. In Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare paper authors have compared three different data mining methods for the prediction of breast cancer survivability. The used ANNs, decision trees and logistic regression. They rank decision tree as the best predictor with 99.6 utilized a large dataset with 10-fold cross validation. Authors made use of C5 algorithm as the decision tree algorithm and computed accuracy, sensitivity and specificity. Decision tree nominated as the best choice for all the above state measures.[4]

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

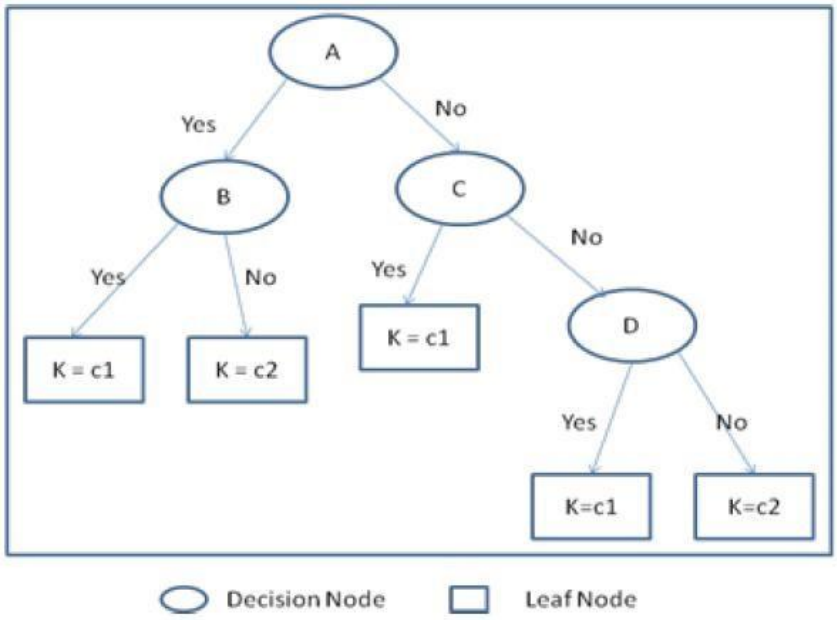


**Fig 3 Decision Tree**

**2.1.2 Artificial Neural Networks**

Artificial neural networks (ANN), also known as neural network attempts to simulate the functional or structural characteristics of biological neural network. ANN should be customized for each application otherwise it may lead to poor performance. Moreover being a black box technology is another disadvantage of ANN. In A neural network model for prognostic prediction paper authors apply ANN on two different breast cancer dataset. Both of these datasets uses the morphometric features. A modified Streets ANN model has been employed. Backpropagation has been used to train the networks. Backpropagation consists of three layers a) input layer b) hidden layer c) output layer. The specified model uses the probability threshold and differentiates the patients with good or bad prognosis. [5] Please note that the probability of recurrence at different time intervals has been predicted to define the probability threshold. Authors have utilized a three-layer feed forwad ANNs with sigmoid activation. Their prediction appears to be good to differentiate good and bad prediction groups. However, for a homogenous group prediction seems to be poor. They claim that their ANN model can help to answer a question like After a surgery, what is the level of recurrence risk for this breast cancer patient compared to all other patients? However, they fail to provide a satisfactory prediction within a more homogenous subgroup.

In Machine learning: a review of classification and combining Techniques paper authors utilize ANN with multilayer perceptron. It uses backpropagation training similar to neural network. Some author uses much bigger dataset and provides a good comparison with the traditional TNM prediction system. They claim that ANN can give much better accuracy as compared to the traditional TNM system. ANNs accuracy reaches as high as 81addition of few more demographic variables.

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

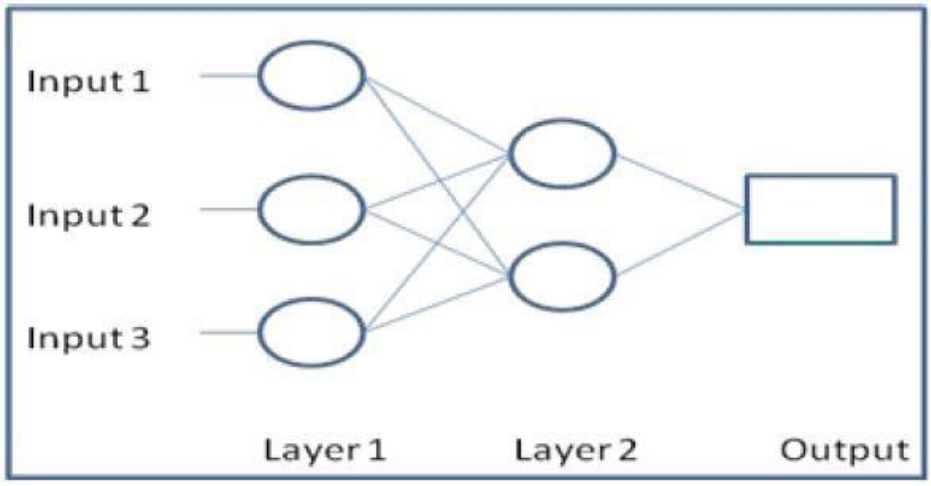


**Fig 4 Artificial Neural Network**

**2.1.3 Support Vector Machine**

Support Vector Machine (SVM) is one in all the supervised machine learning algorithms with well-built regularization properties, that is, the optimisation procedure maximizes prophetic accuracy whereas intentionally it reduces the overfitting of the coaching information. [7]

Basically SVM spins around the idea of finding optimal decision boundaries (maximizing the margin, hence forming largest achievable distance among the separating hyper-plane and the instances on either side of it..

Figure 5 is the example of linear classification of SVM whereas Figure 6 shows the non-linearly separable data. Figure 7, shows the general design to map the original non-linear input feature space to some higher-dimensional linear feature space. C-Support Vector Classification Filter (C-SVCF) algorithm was utilized to spot and eradicate outliers in breast cancer survivability data sets.

Results of their approach indicated performance improvement of breast cancer survivability prediction models by improving data quality. This script distinctively draws attention towards the application of SVM for forecasting better survival rates. Gaussian kernel nonlinear SVM is applied to determine Survival curves and possible chemotherapy decision for new patients by assigning them to one of the three groups (Good Prognosis, Intermediate Prognosis, Poor Prognosis) without the need for lymph node status.

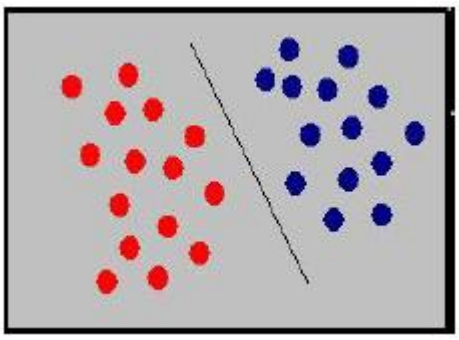

**Fig 5 Linearly Separatable Data Points**

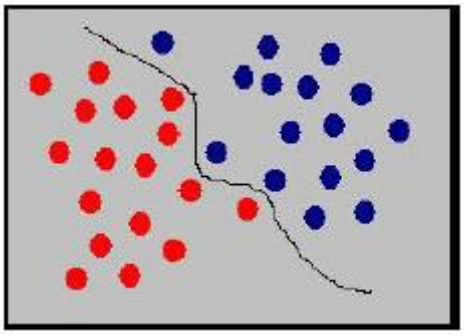

**Fig 6 NonLinearly Separatable Data Point**

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

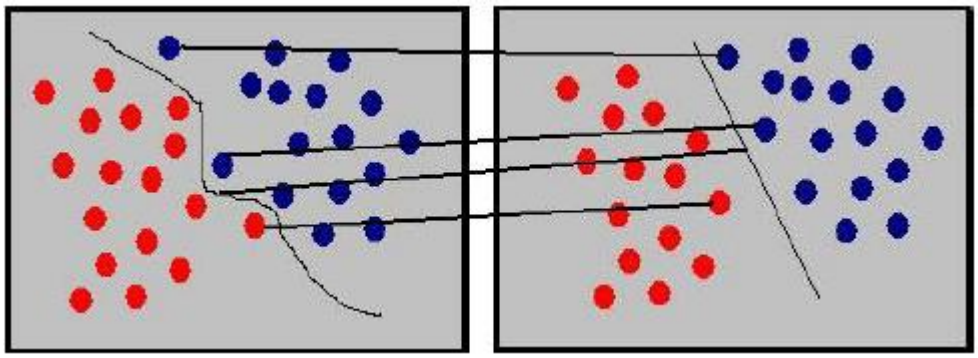

**Fig 7 Non-linear data points separated linearly in high dimensional Feature space**

## III. SURVEY ANALYSIS

This literature survey shows the table in which different techniques, purpose and data sets are defined:

| No | Title of paper | Approach | Data set | purpose |
|---|---|---|---|---|
| 1 | Improved breast cancer prognosis through combination of clinical and genetic markers | I-RELIEF algorithm is used to identify a hybrid signature through combination of both genetic and clinical markers | vant Veers data | predict the likelihood of disease recurrence and metastases in breast cancer. |
| 2 | Artificial Neural Networks Improve Accuracy of Cancer Survival Prediction | TNM staging system, artificial neural networks (ANN) | Cancers breast and colorectal | Comparison of the TNM Staging System with the Artificial Neural network for cancer survival |
| | | | carcinoma Patient Care Evaluation PCE) data | |
| 3 | Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare | Decision Trees | SEER data (1973-2003) is used for breast cancer prognosis | We analyzed the possible potential of fuzzy logic based classifiers, and came up with a conclusion that they can be the natural allies of a physician involved in predictive medicine |
| 4 | Predicting breast cancer survivability: a comparison of three data mining methods | artificial neural networks and decision trees, logistic regression | SEER Cancer Incidence Public-Use Database for the years 19732000 | measure Accuracy, sensitivity and specificity. we defined survival as any incidence of breast cancer where person is still alive after 5 years from the date of diagnosis |
| | | 10-fold cross-validation | | |
| | | preprocessing : data cleansing and transformation | | |
| 5 | wFDT - Weighted Fuzzy Decision Trees for Prognosis of Breast Cancer Survivability | Fuzzy decision tree | SEER data (1973-2003) is used for breast cancer prognosis | analyzed the possible potential of fuzzy logic based classifiers ,and came up with a conclusion that they are fit to act as natural allies of a physician involved in predictive medicine. |

12

**Fig 8 A Survey of Prediction Models for Breast Cancer Survivability**

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

## IV. CONCLUSION

In this paper we have tried to summarize different machine learning and data mining models that are currently being explored for the analysis of breast cancer prognosis. First of all, a brief summary is presented regarding the deadliest impact cancer is having on the world. Secondly, we have tried to summarize the good impact of technology and innovation on the breast cancer identification, treatment and prognosis. We particularly talked about different machine learning methods that are being explored for predicting breast cancer survivability. A growing number of data mining strategies are being used either independently (like ANNs, decision trees etc) or in a hybrid way (like fuzzy logic and decision tree) to improve the prognosis of breast cancer survivability. We have seen that these models have increased the accuracy many fold especially as compared to the traditional statistical based systems. Overall, this paper tends to draw together and review all the latest trends in the field of prediction of breast cancer survivability.

## V. FUTURE SCOPE

So far within the cancer identification domain, there's no general rule for choosing the amount of patterns for the kinds of tumors. Within the future, the simplest way of crucial the amount of symbolic tumors ought to be developed. This may cut back the time for feature extraction. a far better robust an improved applied mathematics analysis of the heterogeneous datasets used would offer more correct results and would offer reasoning to sickness outcomes. Analysis is needed supported the development of a lot of public databases that might collect valid cancer dataset of all patients that are diagnosed with the sickness. Their exploitation by the researchers would facilitate their modeling studies leading to a lot of valid results and integrated clinical higher cognitive process.

## BIBLIOGRAPHY

[1] Bernardi, Luca, et al. "Mining information for functional genomics." *IEEE Intelligent Systems* 17.3 (2002): 66-80.

[2] Sun, Yijun, et al. "Improved breast cancer prognosis through the combination of clinical and genetic markers." *Bioinformatics* 23.1 (2006): 30-37.

[3] Si, Jingna, Jin Cheng, and Rongling Wu. "PFGD: A systematic functional genomics resource for Poplar." *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*. IEEE, 2014.

[4] Yoon, Kihoon, and Stephen Kwek. "An unsupervised learning approach to resolving the data imbalanced issue in supervised learning problems in functional genomics." *Hybrid Intelligent Systems, 2005. HIS'05. Fifth International Conference on*. IEEE, 2005.

[5] Zhang, Min-Ling, and Zhi-Hua Zhou. "Multilabel neural networks with applications to functional genomics and text categorization." *IEEE transactions on Knowledge and Data Engineering* 18.10 (2006): 1338-1351.

[6] Khan, Muhammad Umer, et al. "Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare." *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*. IEEE, 2008.

[7] Su, Hai, et al. "Robust automatic breast cancer staging using a combination of functional genomics and image-omics." *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015.