ISSN: 0974-2115
www.jchps.com Journal of Chemical and Pharmaceutical Sciences

Incorporating Biomarker Test in the Prognosis of Breast Cancer using Random Forest Algorithm

S. Divya Meena*
Department of Computer Science and Engineering, Jansons Institute of Technology, Coimbatore- 641659
***Corresponding author: E-Mail: divyameena.s@jit.ac.in**

ABSTRACT

The universal burden of Breast cancer surpasses all other forms of cancers and the prevalence of breast cancer is increasing dramatically. Breast cancer is the most prominent non-preventable but curable cancer-associated death among women (Tasnuba Jesmin, 2013). The increasing deaths alarm us to screen a healthy woman for the chances of breast cancer, to detect it earlier and in that way lessen the risk of fatal end from this disease. Moreover, earlier detection can help in breast preservation by opting therapeutic option. Considering the national primacies, the focus has been on the cancer aetiology with identification of preventable risk factors, understanding the mechanism of carcinogenesis and predicting the result of the disease before the treatment instigates (Seyyid Ahmed Medjahed, 2013). This is one of the most fascinating yet perplexing tasks, where the data mining techniques can be employed. Therefore, the motive of the paper is to develop a comprehensive healthcare prototype, leveraging the power of Predictive Analytics along with Big Data, to make a shift from Sick-care to wellness-care. The prototype will develop a predictor model for Breast Cancer prognosis (Kawasar Ahmed, 2013). This can be realized by harnessing the massive amount of genomic data churned out by ever-advancing technologies so that they decipher into meaningful cancer prevention and treatment strategies.

KEY WORDS: Breast Cancer**,** Oestrogen Receptor, Progesterone Receptor, Hormone Receptors, Ki67, Chemotherapy.

1. INTRODUCTON

The iconic disease of our time emanates without caution and attacks those who abuse their body and also those who don't; those who have genetic cancer and those who don't. The universe has seen substantial progress in the cancer care quality over the span of years although it still remains one of the most deadly diseases that immediately trigger the fatal end (Kroshnaiah, 2013). With 700,000 deaths, 1.1 million new cases every year, 3.3 million patients at any point in time, Cancer has established itself as the leading killer, blowing out 70% of younger lives. Curing such a deadly disease has been proved to be difficult as Cancer is not a single disease; it's a compound of 100+ diseases that we call cancer. As stated by WHO, Breast cancer is the leading cause of cancer deaths among women all over the world (Ada Ranjneet Kaur, 2013). Though the probability of breast cancer escalates after the 40 years; several aspects like poor lifestyle, higher stress levels, poor diet, irregular sleeping hours, late menopause, early menarche, and increasing maternal age, increases the risk of the breast cancer (Ferlay, 2015). Obesity, smoking and higher alcohol intake is in some cases, a part of the urban woman's lifestyle. All these factors increase the risk of breast cancer by negatively impacting certain hormones and proteins. Breast cancer mortality can be reduced only by detecting it earlier, intervening it and following up on post-operative treatments (Mittra, 2010).

Breast Cancer Scenario in India: Cancer care in India is branded as High incidence, late detection and lack of affordable quality care to the majority of people and so is resulting in high mortality. It is distressing to note that this high percentage of late detection owes to the issues of access, affordability and awareness since both the cost and successful result of the treatment is in favour of earlier detection (Parkin, 2010). It is vital for the patrons of Indian healthcare to address this growing hazard before it turns out to be a national catastrophe. In India, risk prediction in cancer incidence is relatively poor, with late stage diagnosis in common place and a heavy dependence on conventional risk factors such as smoking, BMI and family history (Nagarajan, 2015). Although these factors can altogether predict a reasonable percentage of cancer incidence there is indeed a room for improvement, in particular in early diagnosis which in turn provides greater treatment options and decreased mortality rate. Big initiatives to decipher big data are stepping stones to comprehensive cancer care that integrates cancer genomics into cancer prevention and treatment (Kalaiselvi, 2015).

The primary objective of this study is to develop a predictive model for prophesying the 5-year and 10 year survival rate of cancer patients (Revathy, 2011). The prediction also includes the possibility of survival after taking adjuvant therapy, hormone therapy and chemotherapy. Precise prognosis can help general practitioners to decide whether to initiate a new treatment or continue anti-cancer therapy, in facilitating the shifts to sanatorium care, in enabling appropriate advanced care planning and also in end-of-life decision making.

Research Findings: Ample of work have been done to predict the risk of the cancer. In the research finding, certain traits of our work are compared with previous works. Some of them are as follows;

Table.1. Research findings

Type	Endpoint	Algo	Data	Ref
Breast	Susceptibility	DT	Mixed	Catto
Breast	Recurrence	K-NN	Clinical	Jin
Breast	Survivability	DT	Mixed	Listgarten
Breast	Susceptibility	SVM	Mixed	Delen
Breast	Survivability	Naïve Bayes	Mixed	Bellaachia
Breast	Recurrence	NN	Clinical	Grumett
Breast	Treatment Response	SVM	Clinical	Hayashida
Breast	Recurrence	DT	Clinical	Masic
Breast	Recurrence	ANN	Mixed	Luna
Breast	Survivability	DT	Clinical	Garcia
Breast	Survivability and Treatment response	RF DT	Clinical and Gene data	This paper

Prognosis of Breast Cancer:

Feature Selection: Feature selection is a pre-processing technique that helps in identifying and removing features that are irrelevant to the classification and in so doing produces a reduced data set.

Table.2. Feature selection

Trivial Variables in Breast Cancer	Significant features in Breast Cancer
Race	Age at diagnosis
Marital Status	Progesterone Status
First degree relative with any cancer	Hormone Receptor Status
Second degree relative with any cancer	Ki67 Status
Histologic Type	Oestrogen Status
Primary site of cancer	Tumour grade and size
Breast Biopsy	Chemotherapy

This process will improve the accuracy without altering the relevance of the features. Pre-processing improves the accuracy and so used widely in Healthcare field (Rama Lakshmi, 2013).

Algorithm: Random forest (RF), a random decision forest technique is an ensemble that operates by developing a swarm of decision trees and outputting the mean prediction of the individual trees. It is one of the most popular classification frameworks that can classify large amount of data with accuracy (Kalaiselvi, 2014). They are extensively used in the diagnosis of Breast cancer, endometrial cancer and Heart disease. In the prognosis of Breast cancer, we develop four decision trees each with ER (+/-), HER2 (+/-) and Ki67 (+/-) as the key attributes. The values of ER, HER2 and Ki67 vary with its sign (positive or negative).

Prognosis Procedure: Breast Cancer can be predicted with Biomarker Test that determines the presence of Estrogen Receptor (ER), Progesterone Receptor (PR), Hormone Estrogen Receptor (HER) and Hormone Progesterone Receptor (HPR). Apart from ER, PR and HER2; few other factors also contribute to better prediction. They are Age at diagnosis of cancer, Diagnosis Mode (Screening Based Detection, Symptoms Based Detection), Tumour size, Tumour Grade (Well differentiated, moderately differentiated, and poorly differentiated), No: of positive nodes (lymph nodes), Ki-67 Status and Chemotherapy (Alessandra Saldanha de, 2014).

Breast Cancer Groups: The choice of treatment meant for a cancer patient depends on the Group she belongs to. Normally it is categorized into four groups as;

Table.3. Breast cancer groups

Group	ER	PR	HER2	Treatment
I	+	+	-	Hormone therapy and Chemotherapy
II	+	-	+	Hormone therapy, Chemotherapy and HER2 targeted therapy
III	-	-	+	Chemotherapy and HER2 targeted therapy
IV	-	-	-	Chemotherapy

Breast Cancer Stages: To discover the stage of the cancer, we use the elementary TNM classification system. TNM stands for Tumour size, Nodes ad Metastasis. Tumour size corresponds to the size of the tumour and nodes represent the positive lymph nodes. Metastasis refers to the spreading of cancer to other nearby or distant organs (Dey, 2013). TNM is composed of four stages and each stage describes the size of the tumour and the extent to which it has spread.

Table.4. TNM staging

Stage	Tumour size (T)	Nodes (N)	Metastasis (M)
X	Tumour cannot be evaluated.	Lymph nodes cannot be evaluated	Spread cannot be assessed.
0	No trace of tumour	Cancer not spread to lymph nodes	Cancer has not spread to distant organs
I	Tumour size is <2 cm	Cancer has spread to ipsilateral axillary lymph nodes	Cancer has spread to distant organs
II	Tumour size is 2-5 cm	Cancer has spread to ipsilateral lymph nodes	-
III	Tumour size is >5 cm	Cancer has spread to ipsilateral mammary/supraclavicular lymph nodes	-
IV	Tumour has attached to chest wall and is spread to lymph nodes	-	-

Based on the above table, we find the cancer stage of a patient. The treatment choice (Hormone Therapy, Chemotherapy or any other treatments) depends on the stage of the cancer. The survivability of a patient varies with the treatment.

Breast Cancer Survivability: The table below represents the different factors contributing to breast cancer, their possible values and the allotted score.

Table.5. Breast cancer survivability

Factors	Values	Score	Factors	Values	Score
Age	<40	1	**ER Status**	Positive	1.0
	40-69	2		Negative	0.0
	>=70	3		Unknown	0.5
Node group	0	0	**HER2 Status**	Positive	0.2413
	1	1		Negative	-0.6762
	2-4	2		Unknown	0
	5-9	3	**Ki67 status**	Positive	0.149035
	10-99	4		Negative	-0.1133286
	>=100	5		Unknown	0
Tumour size	0-9	1	**Chemotherapy**	First	1
	10-12	2		Second	2
	20-29	3		Third	3
	30-49	4	**Detection**	Screening	1.0
	>=50	5		Symptoms	0.0
Tumour grade	1	1		Unknown	0.204
	2	2			
	3	3			

With this table, we calculate the survivability of the breast cancer using the following;

ERP = ƥ + ((Þ *0.55) + (ß *0.35) + (ð * 0.84) + (Ŋ *-0.35) + (η * -0.31)) + (Ƨ status * value) + (Ʀ * value))

ERN= ƥ + ((Þ *0.43) + (ß *0.36) + (ð * 0.40) + (Ŋ *-0.15) + (η * -0.20)) + (Ƨ * value) + (Ʀ * value))

Survivability= $(100 * e^{ER}) / (1+ e^{ER})$

Where Þ-lymph nodes, ß- Tumour size, ð- Tumour grade, Ŋ-screening , η- Chemotherapy, Ƨ- Ki67, Ʀ-HER2 value, ƥ- age , ERP- ER positive, ERN- ER negative.

Breast Cancer Recurrence: Cancer recurrence is the possible reappearance of cancer after treatment during which the cancer cannot be detected. The cancer may recur at the same site as before or somewhere near the primary site of the cancer or somewhere far away from the primary site (Van Vliet, 2014). This type of recurrence is called as Local, Regional and Distant recurrence respectively. For predicting the recurrence rate, we use three models as follows;

Ki-67 Inclusive model = 15.31385 + ƙ *1.4055+ Ȥ *(-0.01924) + Ɑ *(-0.02925) + □ *(0 for HER2 negative, 0.77681 for equivocal, 11.58134 for HER2 positive) + □*0.78677 + Ẇ *0.13269.

Ki-67 exclusive model = 18.8042+ ƙ *2.34123 + Ȥ *(-0.03749) + Ɑ *(-0.03065) + □ *(0 for HER2 negative, 1.82921 for equivocal, 11.51378 for HER2 positive) + □*0.04267.

Semi quantitative IHC model= 24.30812+ Ϛ *(-0.02177) + Ɑ *(-0.02884) + □ *(0 for HER2 negative, 1.46495 for equivocal, 12.75525 for HER2 positive) + Ẃ *0.18649.

Where ķ-Nottingham score, Ϛ-ER, Ɑ-PR, □-HER2, Ẃ-KI67, □-Tumour size.

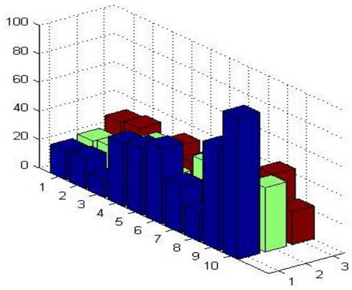

Figure.1. Breast Cancer Recurrence

The chart above represents the different model of recurrence for the same set of values for all there models. The main parameters for recurrence being Nottingham score, ER, PR, HER2 Ki67 and Tumour size.

4. CONCLUSION AND FUTURE ENHANCEMENT

The earlier diagnosis of Breast cancer is a key to effective treatment. In this paper a novel prognosis method in the combination of regression and random forest classifier technique was used to build a breast cancer predictive model. Cancer associated death is increasing intensely. This rate can be reduced only with earlier prediction. But the worst case is that most people avoid cancer screening owing to cost and time associated with it. The model reduces the cost for different medical tests and helps the patients to take precautionary measures well in advance. It leverages the power of both clinical and genomic data to prefigure the cancer risk. It also describes the functional assessment stage of a cancer affected patient. In future this prognosis model can be designed as a web based application and can be implemented in remote areas, to imitate the human diagnostic expertise for predicting the disease. A more efficient model can be built by using different techniques and algorithms. Similar model is to be built for lung and endometrial cancer.

REFERENCES

Ada Ranjneet Kaur, A Study of Lung Cancer Using Data mining Classification Techniques, International Journal of Advanced Research in Computer Science and Software Engineering, 3 (3), 2013, 23-33.

Alessandra Saldanha de, Impact of Diabetes on Cardiovascular Disease, An Update, International Journal of Hypertension, 23, 2013, 234-256.

Dey S, Gupta R, Steinbach M, Kumar V, Integration of Clinical and Genomic data, a Methodological Survey, Technical Report, Department of Computer Science and Engineering University of Minnesota, 2013.

Ferlay J, Cancer Incidence and Mortality Worldwide, IARC Cancer Base No. 11 Lyon, France, International Agency for Research on Cancer, 23, 2015, 123-134.

Kalaiselvi C, Nasira GM, A new approach for the diagnosis of diabetes and cancer using ANFIS, World Congress on computing and communication technologies, WCCCT-IEEE conference, 2014, 188–190.

Kalaiselvi C, Nasira GM, Classification and Prediction of heart disease from diabetes patients using hybrid particle swarm optimization and library support vector machine algorithm, International Journal of Computing Algorithm, 4, 2015, 1403–1407.

Kawasar Ahmed, Tanuba Jesmin, Zamilur Rahman Md, Early Prevention and Detection of Skin Cancer using Data mining, International Journal of Computer Application, 62, 2013, 112-134.

Kroshnaiah V, Narsimha G, Subhash Chandra N, Diagnosis of Lung Cancer Prediction System Using Data mining Classification Techniques, International Journal Of Computer Science And Information Technologies, 4, 2013, 39-45.

Mittra I, A cluster randomized, controlled trial of breast and cervix cancer screening in Mumbai, India, methodology and interim results after three rounds of screening, International Journal of Cancer, 126, 2010, 976–984

Nagarajan S, Chandrasekaran RM, Design and implementation of expert clinical system for diagnosing diabetes using datamining techniques, International Journal of Science and Technology, 8, 2015, 771–776.

Parkin D.M, Boyd L, The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010, Journal of Epidemiol, Community Heal, 29, 2011, 234-256.

Rama Lakshmi K and Prem Kumar S, Utilization of Data mining Techniques for Prediction and Diagnosis of Major Life Threatening Diseases Survivability- Review, International Journal of Scientific & Engineering Research, 14 (6), 2013, 1012-1025.

Revathy N, Amalraj R, Accurate Cancer Classification using Expression of Very few Genes, International Journal of Computer Application, 14 (4), 2011, 1312-1324.

Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi, Abdelkader Benyettou, Breast Cancer Diagnosis using K-Nearest Neibhor with Different Distances And Classification Rules, International Journal of Computer Applications, 45, 2013, 34-42.

Tasnuba Jesmin, Kaswar Ahmed, Badrul Alam Miah Md, Brain Cancer Risk Prediction System Using Data mining, International Journal of Computer Applications, 61, 2013, 45-78.

Van Vliet M.H, Horlings M, Van de Vijver V, Reinders M.J, Wessels L.F, Integration of Clinical and Gene Expression Data Has a Synergetic Effect on Predicting Breast Cancer Outcome, PLoS ONE, 7, 2012.