Analysis of breast cancer event logs using various regression techniques

Saravanan.M.S
Professor, Department of Artificial Intelligence,
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences,
Chennai, India.
saranenadu@gmail.com

Pradnya Patil
Assistant Professor, Department of Computer Engineering, K J Somaiya Institute of Engineering and Information Technology, Mumbai, India.
pradnya08@somaiya.edu

K.Venkata Subbaiah
Professor, Department of Computer Science & Engineering,
DRK Institue of Science and Technology, Hyderabad, India.
kvsubbaiah@gmail.com

2021 International Conference on Computer Communication and Informatics (ICCCI) | 978-1-7281-5875-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICCCI50826.2021.9402360

Abstract — **The breast cancer is a chronic disorder that causes serious illness to the patients despite their age groups. Breast cancer has more number of research to identify the root causes. But in recent research finding also concentrated more on factors affecting the breast cancer with different type of event logs, such as healthcare centers generated data and trail data taken from various webpages. The machine learning techniques are mostly applied on complex type of event logs such as cancer data set, brain tumor dataset and heart related diseases. Among various diseases breast cancer is the one has more complex event logs, which is very complex to analyze and to find the root causes. This research article discuss about the breast cancer data set using logistic regression technique applied with python programming language. This paper also deals about the root causes about the breast cancer and related issues.**

Keywords - Breast cancer, Machine Learning, Healthcare, Event logs, Regression technique.

I. INTRODUCTION

The breast cancer is one of the major research in the recent studies [1], according to the biological research study it has major focus among various serious diseases. Nowadays the survival rate of the breast cancer affected patients are rapidly decreasing [2]. The random forest technique is one which applied on breast cancer dataset in the previous research studies. Machine learning techniques are used to predict good accuracy on breast cancer dataset. The breast cancer has huge number of instances which deals various parameters of disease causes and predictions [3].

Each cell of the human body is analyzed to predict the growth rate or status of Breast cancer to find the normal or abnormal state of the cells [4]. Many number of diagnostic methods are used to identify the cancer cell growth in healthcare diagnostic centers. The radiologist and pathologist are used to diagnose the breast cancer with different tools to prepare the disease history of the patient.

II. RELATED WORK

The breast cancer diagnosis can be done through various machine learning techniques to predict the suspicious cells, whether malignant or benign. The input of data set decides the nature of output depends upon the data processing technique used during the healthcare diagnosis. In breast cancer, the benign is a tumor that does not invade its surrounding tissue or spread around the body [5]. A malignant tumor is a tumor that may invade its surrounding tissue or spread around the body. Cancer cells have the ability to spread to other parts of the body through the blood and lymphatic systems. The popular method to find the nature of the region of interest on breast cancer is by mammography.

Machine learning algorithms can be implemented by two types of learning methods, they are supervised and unsupervised [6]. The supervised machine learning algorithms have test set and training data to predict the target variables, but in unsupervised algorithms do not have the test set to predict the target values. So clustering technique used to predict the unsupervised learning algorithms.

Large number of well labelled images are used for analyzing the breast cancer dataset to predict better results instead of flat files. In most of the cases the machine learning techniques will ensure the quality of prediction by intelligent systems without human intervention, that extract the knowledge such as patterns and rules among the various input values or past experience. The computer aided designing methods are used to automate the breast cancer diagnosis with different intelligent tools. The Support Vector Machine (SVM) algorithm is used to find the large number of features from the input dataset and from this we can differentiate the region of interest in greater extent [7].

In artificial intelligence has lot of opportunities to extract the information from the breast cancer medical images or event logs, using Digital Medical Image Recognition technique we can predict the optimum solution for given dataset [8].

Table 1. Breast Cancer affected rate in India for the year between June'2018 and June' 2019

978-1-7281-5875-4/21/$31.00 ©2021 IEEE

	Men	Women	Total
Population	701,546,980	652,504,878	1,354,051,858
No. of New Cancer Cases	570,045	587,249	1,157,294
No. of Cancer Deaths	413,519	371,302	784,821
5-year prevalent cases	1,000,485	1,257,723	2,258,208
Top 5 most frequent Cancers	1. Lip/Oral Cavity 2. Lung 3. Stomach 4. Colorectal 5. Oesophagus	1. Breast 2. Cervix Uteri 3. Ovary 4. Lip / Oral Cavity 5. Colorectal	1. Breast 2. Lip / Oral Cavity 3. Cervix Uteri 4. Lung 5. Stomach

Particularly in the Asian region the breast cancer is the one which is affected most of the peoples, according to the statistics shown in the above table 1.

The above table has the statistics about the cancer rate of men and women during the year between June' 2018 and June' 2019, hence the rate of cancer affected on breast is top ranked in the third and fourth columns.

III. OBSERVATIONS OF RELATED WORK

The breast cancer event logs with 569 instances taken for experimentation. In this 357 benign and 212 malignant as shown in the figure 1. The malignant are the cells, which spreads the cancer around the tissues and the benign in the other hand has only affected within the cell. The breast cancer event logs has the two types of cells called malignant and benign are shown in the figure 1.

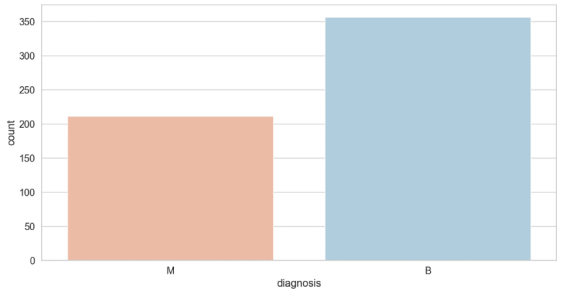

Figure 1. Breast cancer event logs with Malignant and Benign

The benign and malignant are analyzed by converting the event logs in to various parameters they are radius mean, texture mean, perimeter mean, area mean, smoothness mean, compactness mean, concavity mean, concave points mean, symmetry mean, fractal dimension mean.

The figure 2 has some interesting parameters, for example linear parameter between the radius, perimeter and area attributes are suggesting the presence of multiple linear relationship between the various input values.

There are some interesting parameters are showing in the instances that is almost correct linear patterns between the radius, perimeter and area attributes. Another set of values are possibly imply the multiple linear relationship such as concavity, concave points and compactness. The correlations between the variables are shown the scatter plot as in figure 2.

Let's find out if our hypothesis about the multiple linearity has any statistical support.

The patterns shown in the figure 2 are mostly used to represent the plotting of various kinds of benign and malignant cells, which can be represented using various parameters such as radius mean, texture mean, perimeter mean, etc. The parameters are used mostly to represent the physical size of the cells.

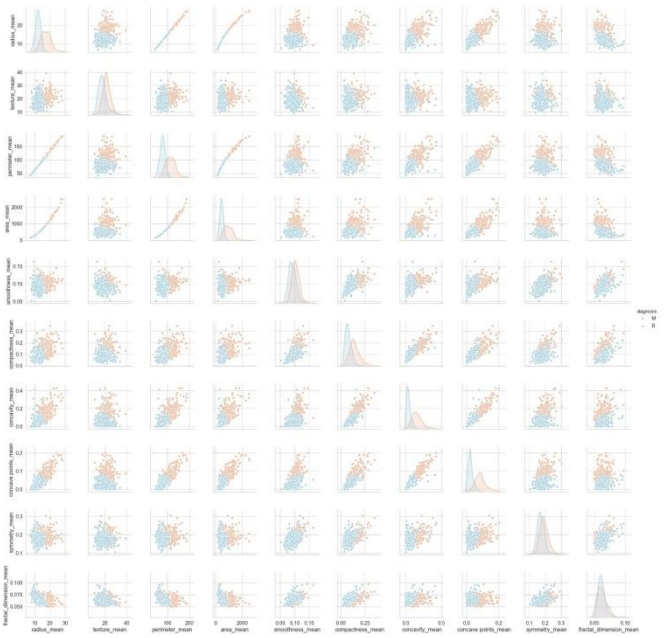

Figure 2. Breast cancer event logs with various parameter representations

Looking at the matrix, we can immediately verify the presence of multicollinearity between some of our variables. For instance, the radius mean column has a correlation of 1 and 0.99 with perimeter mean and area mean columns, respectively. This is probably because the three columns essentially contain the same information, which is the physical size of the observation of the cell. Therefore we should only pick one of the three columns when we go into further analysis.

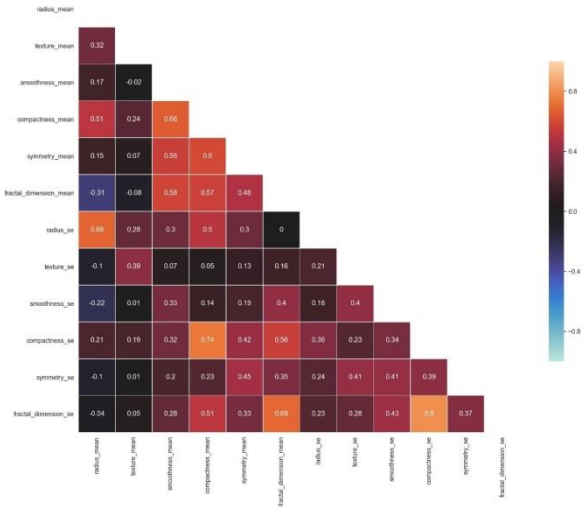

Figure 3. Breast cancer event logs with physical sizes of the cell

IV. BREAST CANCER EVENT LOG ANALYSIS USING LOGISTIC REGRESSION

To develop the model using logistic regression first we have to begin by splitting the dataset into two parts, one as a training set for the model, and the other as a test set to validate the predictions that the model will make. If we omit this step, the model will be trained and tested on the same dataset, and it will underestimate the true error rate, a phenomenon known as over fitting.

The figure 4 shows the number of breast cancer cells labelled with percentage. This percentage shows the input ratio of overall test set and training set. The various patterns such as radius mean, texture mean, smoothness mean, compactness mean, symmetry mean, fractal and dimension mean.

```
Number of cells labeled Benign:  357
Number of cells labeled Malignant :  212

% of cells labeled Benign 62.74 %
% of cells labeled Malignant 37.26 %
diagnosis ~ radius_mean + texture_mean + smoothness_mean + compactness_mean + symm
etry_mean + fractal_dimension_mean + radius_se + texture_se + smoothness_se + comp
actness_se + symmetry_se + fractal_dimension_se
```

Figure 4. Number of Breast cancer cells labeled with percentage

It is like writing an exam after taking a look at the questions and answers beforehand. We want to make sure that our model truly has predictive power and is able to accurately label unseen data. We will set the test size to 0.3, that is 70% of the data will be assigned to the training set, and the remaining 30% will be used as a test set. In order to obtain consistent results, we will set the random state parameter to a value of 40. Now that we have split our data into appropriate sets, let's write down the formula to be used for the logistic regression.

The formula includes all of the variables that were finally selected at the end of the previous section. We will now run the logistic regression with this formula and take a look at the results.

To find the final model first we will feed in the test data to this model to get predictions of values. Then, we will evaluate how accurately the model have predicted the data. This model can take some unlabeled data and effectively assign each observation a probability ranging from 0 to 1. This is the key feature of a logistic regression model. However, for us to evaluate whether the predictions are accurate, the predictions must be encoded so that each instance can be compared directly with the labels in the test data.

In other words, instead of numbers between 0 or 1, the predictions should show "M" or "B", denoting malignant and benign respectively. In our model, a probability of 1 corresponds to the "Benign" class, whereas a probability of 0 corresponds to the "Malignant" class. Therefore, we can apply a threshold value of 0.5 to our predictions, assigning all values closer to 0 a label of "M" and assigning all values closer to 1 a label of "B".

If this is confusing, let's go through this step-by-step. We can confirm that probabilities closer to 0 have been labeled as "M", while the ones closer to 1 have been labeled as "B". Now we are able to evaluate the accuracy of our predictions by checking out the classification report and the confusion matrix, it is shown in the figure 5.

Our model have accurately labeled 96.5% of the test data. This is just the beginning however. We could try to increase the accuracy even higher by using a different algorithm other than the logistic regression, or try our model with different set of variables. There are definitely many more things that could be done to modify our model, but I will conclude this report here for now.

```
              precision    recall  f1-score   support

           B       0.982     0.965     0.974       115
           M       0.931     0.964     0.947        56

    accuracy                          0.965       171
   macro avg       0.957     0.965     0.961       171
weighted avg       0.966     0.965     0.965       171

Confusion Matrix:
 [[111   4]
 [  2  54]]

True Negative: 111
False Positive: 4
False Negative: 2
True Positive: 54
Correct Predictions 96.5 %
```

Figure 5. Breast Cancer various parameters of such as accuracy, recall, fi-score, support, Confusion matrix with percentage of correct prediction.

V. CONCLUSION

This paper conclude that the breast cancer disease is a one which causes serious illness to the patients despite their age groups. In the previous research studies the Breast cancer has contributed more in many research findings. But this paper discusses the new event log with 569 instances and also discussed more on factors affecting the breast cancer with different type of event logs, such as healthcare centers generated data and trail data taken from various webpages. The machine learning techniques are mostly applied on complex type of event logs such as cancer data set, brain tumor dataset and heart related diseases. Among various diseases breast cancer is the one has more complex event logs, which is very complex to analyze and to find the root causes. Therefore this research article discuss about the breast cancer data set using logistic regression technique applied on python programming language. In future this paper gives a research gap to identify the early stage event logs of breast cancer and related issues.

REFERENCES

[1] D. S. Jacob, R. Viswan, V. Manju, L. PadmaSuresh and S. Raj, "A Survey on Breast Cancer Prediction Using Data MiningTechniques", 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), Tiruchengode, 2018, pp. 256-258 doi: 10.1109/ICEDSS.2018.8544268.

[2] D. Srinivasan, M. Gopalakrishnan and D. Srinivasan, "A survey on antennas used for early detection of breast cancer," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, 2017, pp. 1736-1739, doi: 10.1109/ICECDS.2017.8389746.

[3] B. K. Gayathri and P. Raajan, "A survey of breast cancer detection based on image segmentation techniques," 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), Kovilpatti, 2016, pp. 1-5. doi: 10.1109/ICCTIDE.2016.7725345.

[4] T. Padhi and P. Kumar, "Breast Cancer Analysis Using WEKA", 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 229-232. doi: 10.1109/CONFLUENCE.2019.8776911.

[5] A. Li et al., "Association Rule-Based Breast Cancer Prevention and Control System," in IEEE Transactions on Computational Social Systems, vol. 6, no. 5, pp. 1106-1114, Oct. 2019. doi: 10.1109/TCSS.2019.2912629.

[6] Saravanan.M.S, R. Rajpriya, Survey Paper on Cervical Cancer Detection Through Artificial Intelligence Techniques, Published in International Journal of Computer Trends and Technology by Seventh Sense Research Group publishers, India, Volume 57, Number 2, March 2018, pp. 98-101, ISSN: 2231-2803.

[7] Ashis Jalote-Parmar, Petra Badke-Schaub, Wajid Ali, Eigil Samset", Cognitive processes as integrative component for developing expert decision-making systems: A workflow centered framework", Journal of Biomedical Informatics, Vol. 43, pp. 60–74, 2010.

[8] Saravanan.M.S, "Breast Cancer Classification Using Visual Data Mining Techniques", Published in the International Journal of Advanced Science and Technology by SERSC Publications, India, Vol. 28, No. 13, December 2019, pp. 188-197, ISSN: 2005-4238.

[9] Shaikh Afroz Fatima, "Biological Early Brain Cancer Detection Using Artificial Neural Network", International Journal of Computer Science and Network (IJCSN), Vol. 1, No. 4, pp. 11-15, 2012, www.ijcsn.org ISSN 2277-5420.

[10] Arati Kothari, "Detection and classification of brain cancer using artificial neural network in MRI images", World Journal of Science and Technology, Vol. 2, No. 5, pp.1-4, 2012, ISSN: 2231 – 2587..