International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

Breast Cancer L and R Classification and Analysis using Machine Learning Techniques

Ravinthiran Partheepan
Faculty of Informatics, Master Programme – Applied Informatics, Vytautas Magnus University, Kaunas, Lithuania

Abstract:* This paper represents the integration of machine learning techniques such as random Forest, KNN, Neural Network, Naaive Bayes, Logistic regression and Classification tree. The implementation of breast cancer dataset could be used for the predictive measures of test accuracy, classification and their sensitive values. Breast cancer has been a major cause for the death of womens now a days. This disease has no age limits, it could affects any aged women. If the breast cancer is positive which means if a women is affected by this type of cancer the ratio of survival are very low but if it has been detected at the earlier stage the chances for the survival are very high. The two major roles takes place in diagnosing the breast cancer are malignant and benign level. These roles states that the benign level is low and the malignant level is high. This paper reports the accuracy of diagonising and the predictive measures of several outcomes such as recurrence and non-recurrence events totally classified as tumor, L and R breast cancer, cancer affected on which age durability and it's classification whether the cancer affected on left or right breast. Many existing works are available but they didn't provide the accuracy level of the outcomes.

***Index Terms:* Breast Cancer, Random Forest, KNN, Neural Network, Naive Bayes, Logistic Regression, Classification Tree, Machine Learning Approaches.**

I. INTRODUCTION

Machine learning shows a rapid interest in this last decade. This machine learning core could integrates low cost memory and cheaper computing power. An enormous of data could be processed and analyzed effectively. The machine learning plays a major role in several applications such as natural language processing, expert system, Image recognition and data mining and it's prediction. This research provides the core of accuracy of several machine learning method on breast cancer diagnosis which is L and R classification.

The breast cancer occurs mostly for the 40 years aged women and above , when the cells in the glands that produce milk are divide drastically and abnormal. The breast cancer becomes the most common form of cancer affecting women. Medical experts determines the recommended biopsy and mammogram. This paper demonstrates various machine learning methods with the workflow for breast cancer diagnosing.

This research contributes the machine learning approaches which includes random forest, K-nearest neighbour, neural network, naive bayes, logistic regression and classification tree. The data imported in this research work is the diagnostic breast cancer. On the second second hand contribution the feature extraction and feature selection technique will be invoked. Now a days, 90% of womens are affected and in that 45% are dead due to this breast cancer. So this classification methodology could provides accuracy & test score of diagnosis.

II. RELATED WORKS

A. Rupali R. Gangande, Desta Mulatu used a data mining techniques for prediction of breast cancer recurrence approach. In this system they have used a classification association rules, methods and formulated the results on diagnosis.

B. C.Raghavendra, K. Sai saranya, K. Rajendra Prasad has detailed the application on classification of breast cancer detection using feature extraction technique . In this method with the usage of extracted feature for cancer dataset and provides the outcome on support vector machine learning.

C. Sornxayya, Phetlasy has proposed a data classification with hybrid method for how different classifier algorithms. The classifier first measures to reduce false negative classifier and on the second measures to reduce false positive. Algorithms such as naive bayes(NB),decision tress J48, sequential minimal optimization(smo) for support vector machine, Instance based learning Ibk algorithm for K-nearest neighbour and multilayer perception(MLP) for neural network.

D. Shahed anzarus sababa, Md. Ahadur rahman munsi, Shihabuzzaman shihab, Ahmed Iqbal predicts whether the breast cancer is occurred in recurrence or non-recurrence event. Implemented on C4.5 decision tree, Support vector machine and naive bayes classification algorithms.

III. PROPOSED NOTION

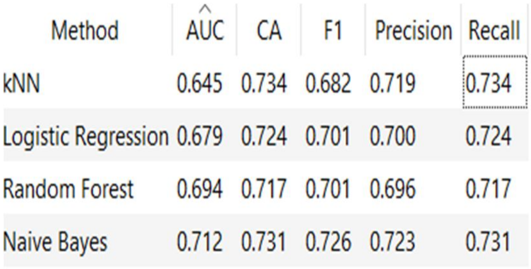
This application system scopes in formulating the enriched accuracy results of various machine learning approaches in classification and prediction of breast cancer diagnosis. The data can be tested with two major classification such as feature selection and feature extraction. By the observation of the predicted results, it could be classified with precision and recall of the recurrence and non-recurrence events. By the measures of the precision and recall the overall evaluation would be compared with several approaches K-nearest neighbour, Random forest, Neural network, naive bayes and logistic regression to predict out the best accuracy of overall methods and the accuracy on each methods to define which method is efficient for classification and analysis of breast cancer diagnosis.

IV. PROPOSED APPROACHES

These following machine learning algorithms are composed in formulating the accuracy test on each approaches for breast cancer diagnosis such as,

A. K-Nearest Neighbor
B. Random forest
C. Naive Bayes
D. Logistic Regression

The classifications on diagnosis could be checked with the classification tree technique.



Method	AUC	CA	F1	Precision	Recall
kNN	0.645	0.734	0.682	0.719	0.734
Logistic Regression	0.679	0.724	0.701	0.700	0.724
Random Forest	0.694	0.717	0.701	0.696	0.717
Naive Bayes	0.712	0.731	0.726	0.723	0.731

Fig.1 Overall Accuracy report on Machine Learning for Breast cancer Diagnosis

1) *K-Nearest Neighbor(KNN):* KNN approach is type of lazy learning, such that it takes local approximation into the core and all computation might be referred in some cases until the function evaluation. In K-nearest neighbour both classification and regression can be used to assign the weight of the neighbour, by this case that the nearest neighbour could contribute the focus on more to the average than the more distinct values. In classification phase, the k is user defined constant and the most frequent assigned among the k training node is contributed to unlabelled vector. The major two metrices which are used in distance measurement such terms are continous variables also called as eucledian distance. The classification accuracy of K-nearest neighbour could be leveraged only if the specialized approach is learned is distance metric such as neighbourhood component analysis. The K-nearest neighbour classifier could be analyzed in terms of weight = 1/k and all others 0 weight.

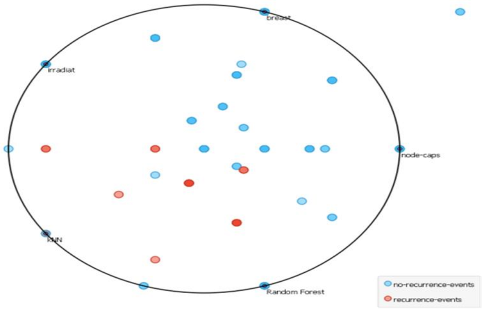
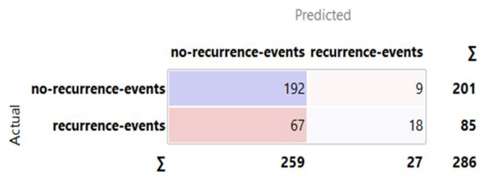


Fig. 2 K-NN visualization report on Breast cancer diagnosis with respect to recurrence and non recurrence event.

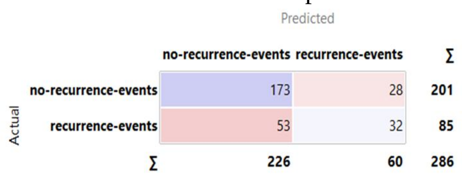


Predicted

Actual		no-recurrence-events	recurrence-events	Σ
	no-recurrence-events	192	9	201
	recurrence-events	67	18	85
	Σ	259	27	286

Fig.3 Confusion Matrix report on recurrence and non recurrence event on breast cancer diagnosis with respect to K-NN

2) *Random Forest:* Random Forest consists of a enormous number of individual entities of decision trees that operate as an ensemble. The core notion behind random forest is simple but it is an effective approach. In random forest each decision tree opt out a class with maximum entity of node or value and a class prediction.



Predicted

Actual		no-recurrence-events	recurrence-events	Σ
	no-recurrence-events	173	28	201
	recurrence-events	53	32	85
	Σ	226	60	286

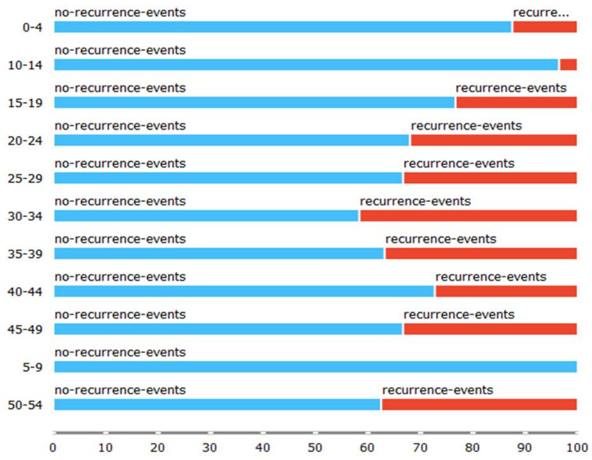
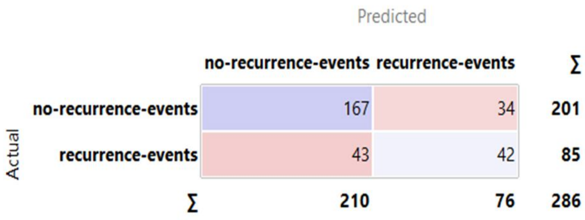


Fig.4 Distribution report on tumor size with the analysis of random forest technique

3) *Naive Bayes Algorithm:* It is a technique that helps to formulate classifiers. The classifiers would classify the objective instance and provide them class labels which are contributed as feature value or vectors of predictors. The naive bayes is based on bayes theorem and the value of a feature is not dependent of the other feature. It states that the moderating the feature value will not affect the value of other features. This algorithm works efficiently for enormous datasets. It could be calculated using the priors and posterior probability of class such as P(recurrence | non recurrence). This could be evaluated as a terms P(recurrence | non recurrence) = (P(recurrence | non recurrence) * P(recurrence) / P(non recurrence)).



Predicted

Actual		no-recurrence-events	recurrence-events	Σ
	no-recurrence-events	167	34	201
	recurrence-events	43	42	85
	Σ	210	76	286

Fig.5 Confusion matrix for recurrence and non recurrence event on breast cancer diagnosis with respect to naive bayes algorithm.

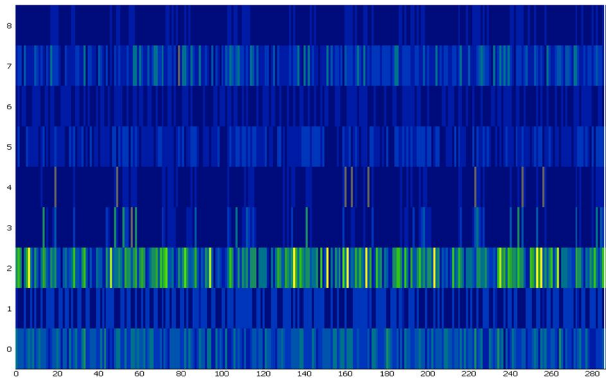


Fig.6 Spectral Visualization on breast cancer diagnosis.

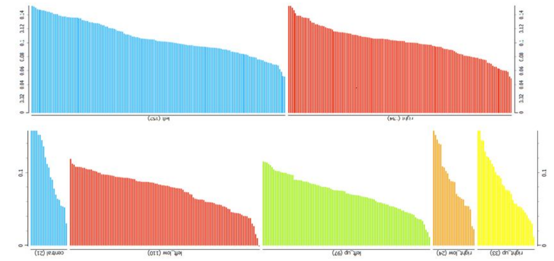
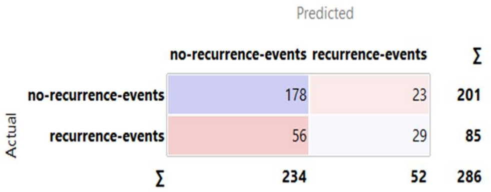


Fig.7 Silhouette report on tumor size and classification breast cancer such as left or right.

4) *Logistic Regression:* It is one of a classification technique in which if the decision threshold is dependent on the classification problem then the setting of threshold is an important aspect in classification core. The decision value of threshold is mostly affected value of recall and precision. The threshold value integrates precision and recall tradeoff such can be classified as low precision/high recall and high precision/low recall. In low precision/high recall function could reduce the number of false negative without reducing the number of false positive. In other hand, high precision/low recall function could reduce the number of false positive without reducing the number of false negatives.



Predicted

Actual	no-recurrence-events	recurrence-events	∑
no-recurrence-events	178	23	201
recurrence-events	56	29	85
∑	234	52	286

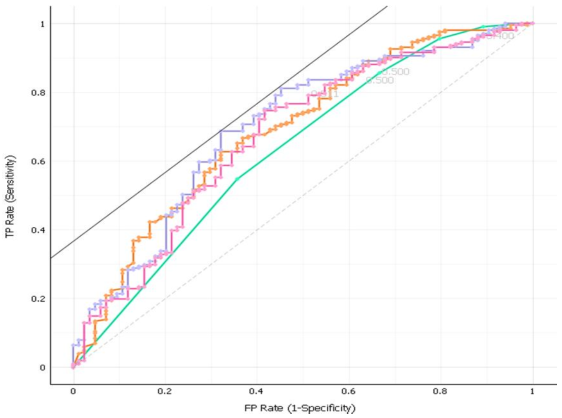


Fig.8 ROC analysis on breast cancer diagnosis non recurrence event.

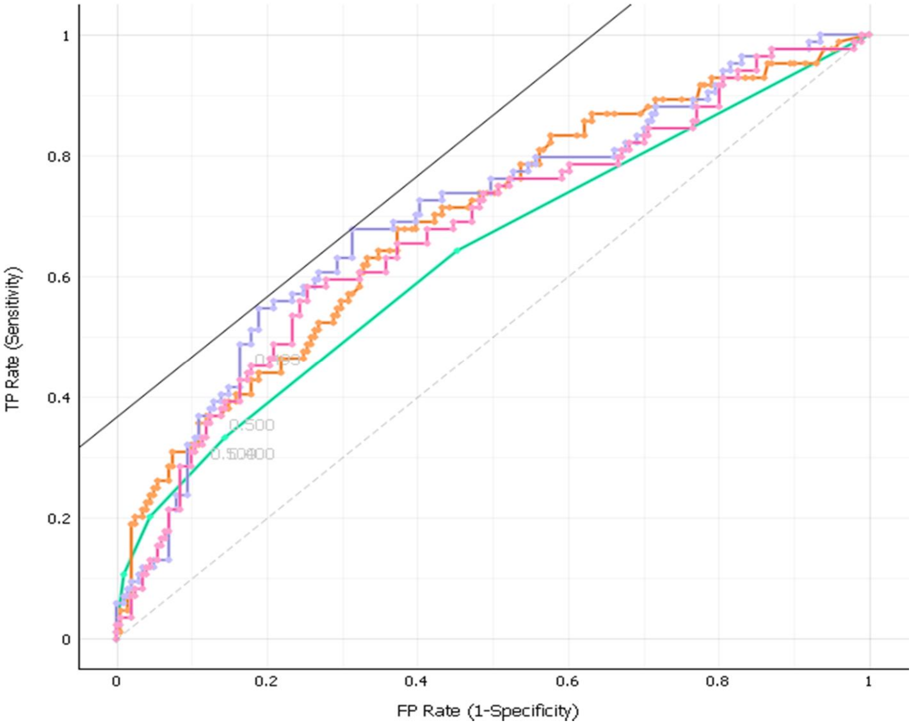


Fig.9 ROC analysis on breast cancer diagnosis recurrence event.

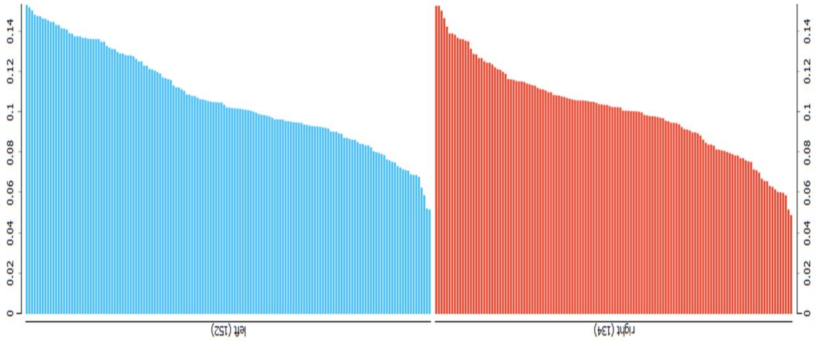


Fig 10. Silhouette report cancer affected breast whether it's Left or Right

V. MERITS AND DE-MERITS

A. Merit

1) This system could provide better accuracy with the comparison of various machine learning approach.
2) Depends on higher accuracy of various machine learing approaches, it is quite easy to classify the breast cancer diagnostics.

B. De-Merit

Test score could take some time to classify the individual technique diagnositic accuracy report which depends on data.

VI. CONCLUSION

In this proposed application, I have formulated the report on various machine learning techniques to leverage the accuracy on breast cancer diagnosis. This application has to be tested with the various machine learning algorithms such as neural network, logistic regression, naive bayes, random forest and k-nearest neighbour. These methods of classification through machine learning algorithm is effective, reliable and could nbe used for any kind of classification on different available datasets. It reports the proliferated accuracy of the system based on the machine learning approaches.

REFERENCES

[1] Ashish Agarwal, Eugene Brevdo, Martin Abadi, Paul Barham, Craig Citro, Zhifeng Chen, Andy, Davis, Greg S.Corrado, Matthieu Devin, Jeffrey Dean, Ian Goodfellow, Sanjay Ghemawat, Geoffrey Irving, Andrew Harp, Yangqing Jia, Michael Isard, Lukasz Kaiser, Rafal Jozefowicz, Josh Levenberg, Manjunath Kudlur, Rajat Monga, Dan Mane, Derek Murray, Sherry Moore, Mike Schuster, Chris Olah, Benoit Steiner, Jonathon Shlens, Kunal Talwar, Ilya Sutskever, Vincent Vanhoucke, Paul Tucker, Fernanda Viegas, Vijay Vasudevan, Pete Warden, Oriol Vinyals, Martin Wattenberg, Yuan Yu, Xiaoqiang Zheng and Martin Wicke. 2015. Tensorflow: Large Scale Machine Learning on heterogeneoussystems.(2015)http://tensortflow.org/ Software available from tensorflow.org.

[2] Yang yuanyuan, Shen Runjie and Shao Fengfeng 2014 Intelligent breast cancer prediction model using data mining techniques proc.in 2014 6th Int .Conf. on Intelligent Human-Machine Systems and Cybernetics 384-387

[3] Ibrahim Roliana, Selamat Ali, Nematzadeh Zahra 2015 Proc.in 2015 10th Asian Control Conf.(ASCC) (IEEE) Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques 1-6.

[4] Tahir N.M. and Hasan H. 2010 “Feature selection of breast cancer based on principal component analysis”, in Signal Processing and Its Applications (CSPA) 2010 6th International Colloquium on 1-4.

[5] Shadgar Bita and Osareh Alireza machine learning techniques to diagnose breast cancer 2010 5th Int. Symp. On health Informatics and Bioinformatics (Antalya, Turkey, 20-23 April 2010)

[6] Sornxzayya, Phetlasy, et al. “Sequential Combination of two Classifier Algorithms for binary Classification to improve the accuracy.” 2015 Third International Symposium on Computing and Networking (CANDAR). IEEE, 2015.

[7] Rupali R. Gangarde, Desta Mulatu. “Survey of Data Mining Techniques for Prediction of Breast Cancer Recurrence”, International Journal of Computer Science and Information Technologies, Vol.8 (6), 2017, 599-601.

[8] Rangaan, R.M., Desautels, J.E.L., Shen, L.:Application of shape analysis to mammographic classifications. IEEE Trans. Medd. Imag. 13(2), 263-274 (1994).

[9] Prabhakar Radhika and Pandey Poonam 22016 2016 1st India Int. Conf. On Information Processing (IICIP) (IEEE) An analysis on machine learning techniques (J48 and AdaBoost) – for classification 1-6 India.

[10] SteinbatchM., Introduction to Data Mining