THINK INDIA JOURNAL

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

Breast Cancer Predication Using Machine Learning Methodologies

Reetu Malhotra[a], Gittaly[b]

[a] Chitkara University Institute of Engineering and Technology, Chitakara University, Punjab, Punjab, India
reetu.malhotra@chitkarauniversity.edu.in
[b] School of Electronics and Electrical Engineering, Lovely Professional University, Phagwara, Punjab, India
gittaly.25304@lpu.co.in

Abstract -In worldwide, the most common cancer amid women is breast cancer. Recently, breast cancer is become the popular source of death in both undeveloped and developed countries and its accountability is increasing every year. Early finding of diseases can be treatable with a diminutive amount of human efforts. Recent development in digital image processing techniques and computer vision are considerably explored in medical applications. These techniques lead pathologists to find cancer in an efficient manner and works as a second opinion for doctor. This article utilized innovative classification models to classify breast cancer diseases. Five classifiers such as Support Vector Machine (SVM), Random Forest (RF), Decision Tree, XGBoost and Neural Networks are used for cataloging of breast cancer.

Keywords- Classification, breast cancer dataset, Feature extraction Machine learning methods.

1. Introduction

Cancer is a life-threatening disease. People lives can save by proper treatment of cancer. Now days, breast cancer become the prime cause of death of women. After lung cancer, it is most dangerous cancer. World Cancer Research Fund provides statistics that in year 2018, around 626,679 [1] deaths and more than 2millions new cancers cases were estimated. Due to the structure of humanoid body, females are more susceptible to breast cancer than men. Age, breast density, family history, alcohol intake and obesity can be the reasons for breast cancer [2-3]. Machine learning based classification systems demand is increasing gradually in medical diagnosis. The assessment of data collected from the patient and expert's decision are the most significant features in diagnosis. Machine learning methodologies, also help medical experts to find possible errors. Because of fatigued or

Copyright © 2019 Author

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

inexperienced experts sometimes errors occur, that can be avoided or minimized in a short time by classification system [4].

1.1 Requirement of cancer detection

Statistics disclosed that in the previous years, condition has become more worse [5]. Appropriate treatment of breast cancer protects people's lives. Identification and detection of the cancer tissues (malignant and benign) is a noteworthy step for treatment of cancer. With the progressive modern photography methodology, targeted (cancer affected) part of human body can be captured more reliably.

1.2 Diseases analysis platform

Machine learning offers an approach for developing automatic, and objective algorithms for analysis of high-dimensional data [6]. Machine learning models have been applied to the training dataset along with extracted features to calculate the testing dataset with input parameters availability. Fig. 1 exhibits the machine learning methodology.

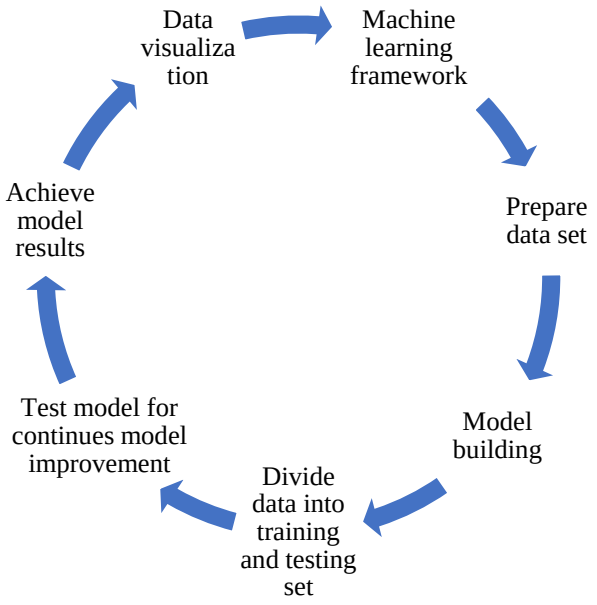

Figure 1: Machine learning methodology

In this proposed study, 5 classification models are used to analyse the breast cancer. Using these classifiers accuracies are measured for 70-30% training and testing partition. Region of convergence, specificity, sensitivity matrices analysis is used to verify performance of classifier. The paper is arranged as listed below. Section 2 defines the background information

Copyright © 2019 Author

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

regarding breast cancer and its problems. Proposed work is briefly defining in section 3. In experimental outcomes are defined in section 4. Finally, in section 5 conclusion is described.

2. Methodology

Using Wisconsin Diagnostic Breast Cancer (WDBC) dataset, five machine learning models are trained to identify cancer. In the dataset, 569 points are considered, where 212 data samples are malignant and 357 data samples are benign. The extracted features are radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, concave points and fractal dimensions. These all features contain three data information as standard error, mean and mean of largest values. So, total dataset contains 30 features. Figure 2. Represents the proposed methodology.

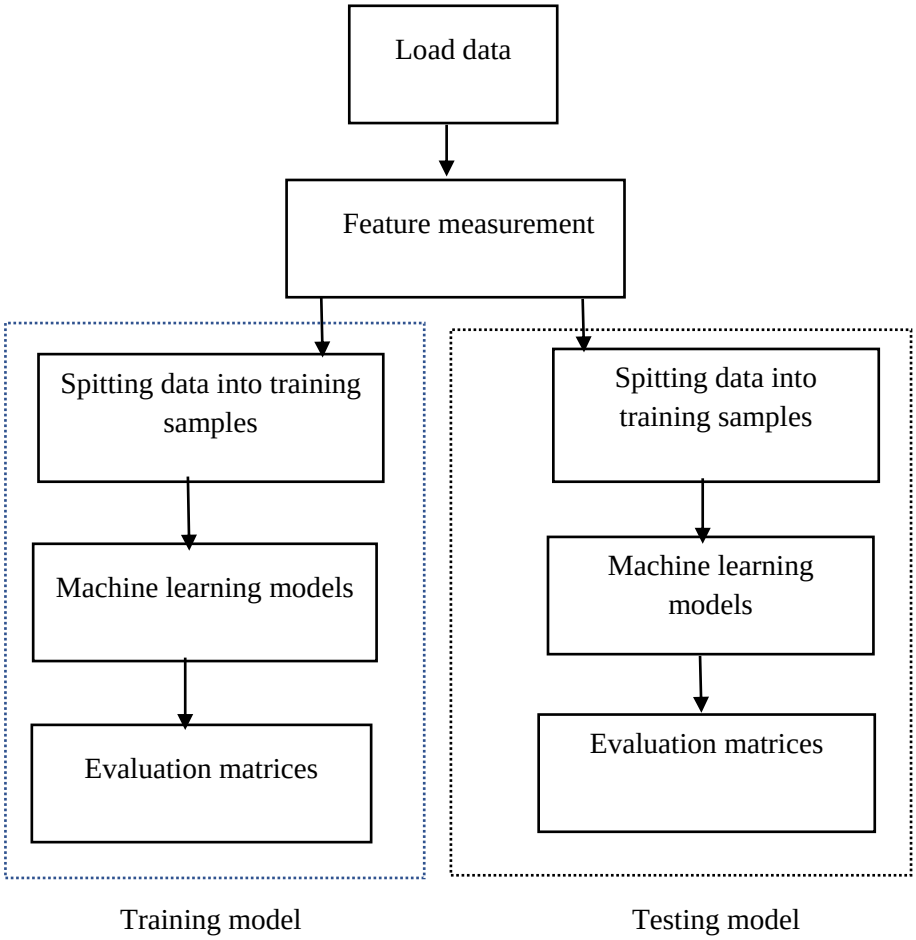

Figure 2: System model

2.1 Machine learning methods

Copyright © 2019 Author

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

This segment represents the machine learning algorithms decision tree, random forest, XGBoost, neural networks and SVM used in the study. The brief description of all models is listed below.

2.1.1 Decision tree

A Decision tree is representing as tree structure. Internal node individually denotes a test on a feature and every division signifies a consequence of test, and every terminal node embraces an output label [7-8]. A tree learned by dividing the basis set into different subsets using attribute value test. Process is recurring on every resulting subset in a recursive way is called recursive partitioning. When all the target variables have same value that time recursion process stops.

2.1.2 Random forest

Ensemble Classifier that develops from decision trees. To categorize a new case, every decision tree delivers a cataloguing label for each input data, then assembles classification labels and selects maximum voted estimate as final output results [9]. Each tree input is sampled from the unique dataset. Features subset is randomly nominated from the features to produce at each individual node [10]. Every tree is grownup without pruning. Random forest is ensemble weak classifiers to create a strong structure classifier.

2.1.3 Neural networks

A sequence of algorithms that endeavors to identify fundamental associations in a group of data through a procedure that lookalikes the method the human brain behaves. It refers to neurons, either artificial or organic in nature [11]. Neural networks can adjust to altering input; so, network produces the highest possible outcome without requiring to reform the output criteria.

2.1.4 XGBoost

XGB is a high-performance execution of gradient boosted decision trees [12]. It trains models in sequence rather than training all models and modified errors made by the previous ones. Models are added successively until no additional developments can be made.

2.1.5 Support Vector Machine

Copyright © 2019 Author

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

Discriminative classifier distinct by a unscrambling hyperplane with specified labeled training data and procedure outputs an optimum hyperplane that classifies new examples. This hyperplane in a two-dimensional plane is a line separating a plane in two segments [13]. It is used both in regression and classification problems with different kernel functions.

3. Results

All experimentations are conducted on a laptop computer using Matlab (2018) and R open (version 3.2.2) software on HP, 2.2-GHz platform. Table 1 demonstrates the manually given hyper-parameters used for the machine learning algorithms. Figure (3), (4) and (5) represents the results of proposed system with respect to cost curve, predicted vs observed values and accuracy [14].

$$\text{Precision (Positive predicted value)} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positives}} \tag{1}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive } + \text{False Negative}} \tag{2}$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negatives}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}} \tag{3}$$

$$\text{Specificity(True neagtive rate)} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{4}$$

$$\text{Negative predicted valve (NPV)} = \frac{\text{False Positive}}{\text{True Negative} + \text{False Positive}} \tag{5}$$

Cost curve demarcated classifiers presentation using cost of misclassification. The x-axis signifies likelihood cost function and y-axis characterizes standardized predictable misclassification cost [15].

$$\text{x axis} = \text{PC}(+) = \frac{\text{p}(+) * \text{C}(-|+)}{\text{p}(+) * \text{C}(-|+) + \text{p}(-) * \text{C}(+|-)}$$

Copyright © 2019 Author

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

(6)

$$\text{y axis} = (1 - \text{PD} - \text{PF}) * \text{PC}(+) + \text{PF} \tag{7}$$

In all the experimentation, Random Forest performs best with accuracy 97.89% as compares to others.

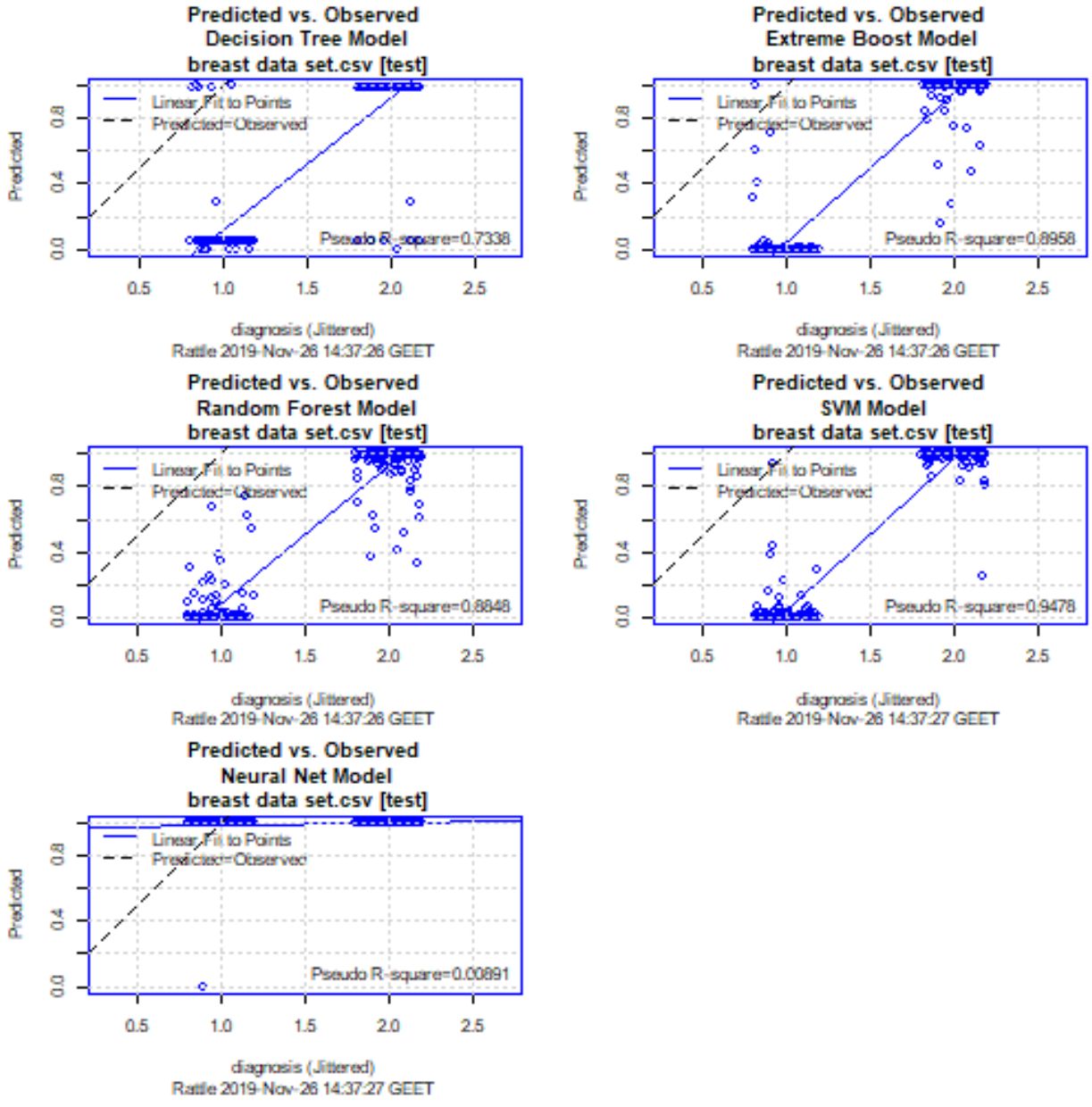

Figure 3: Comparison analysis of Predicted vs observed values of classification models

Copyright © 2019 Author

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

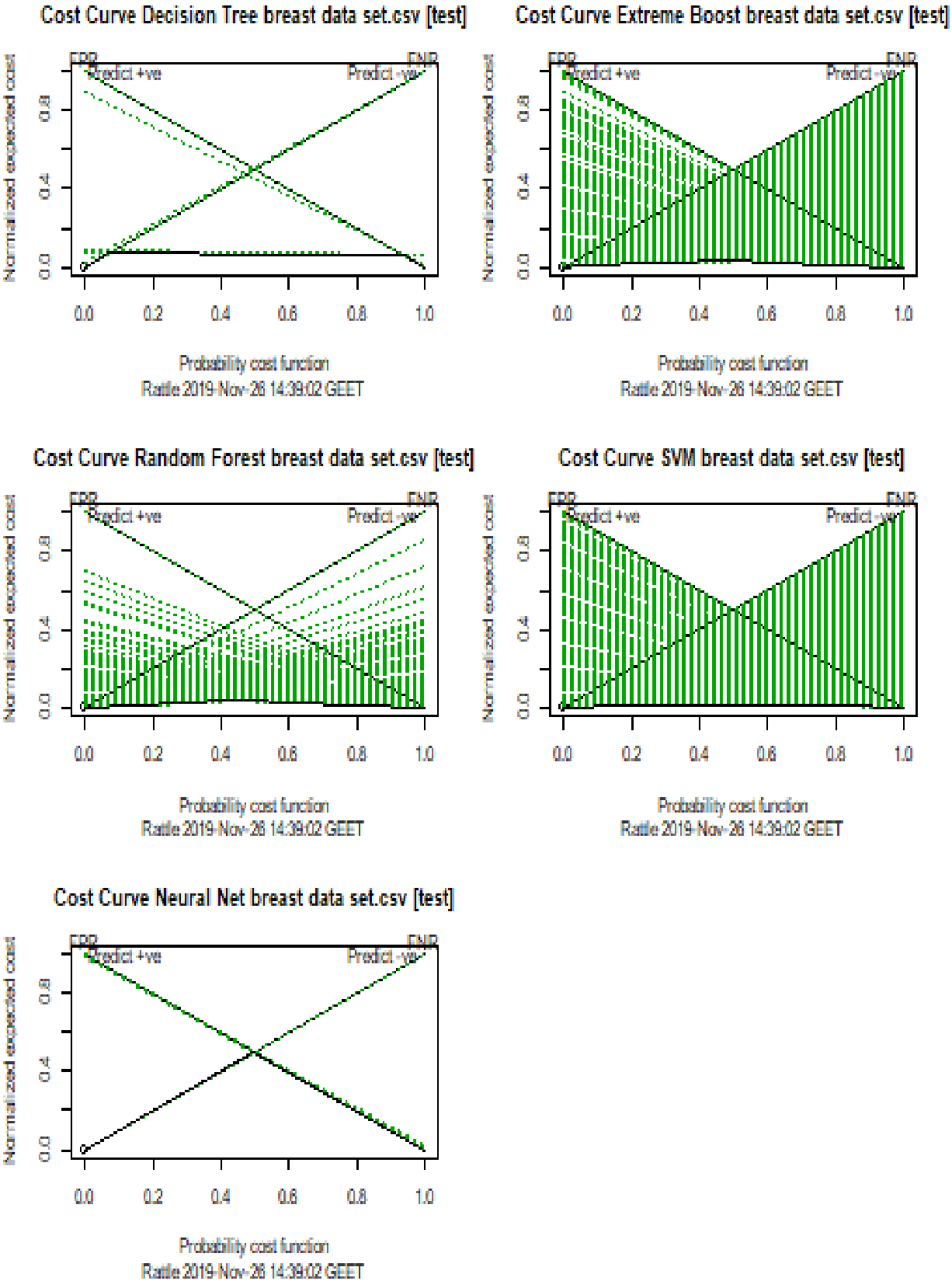

Figure 4: Comparison analysis of cost curve values of classification models

Copyright © 2019 Author

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

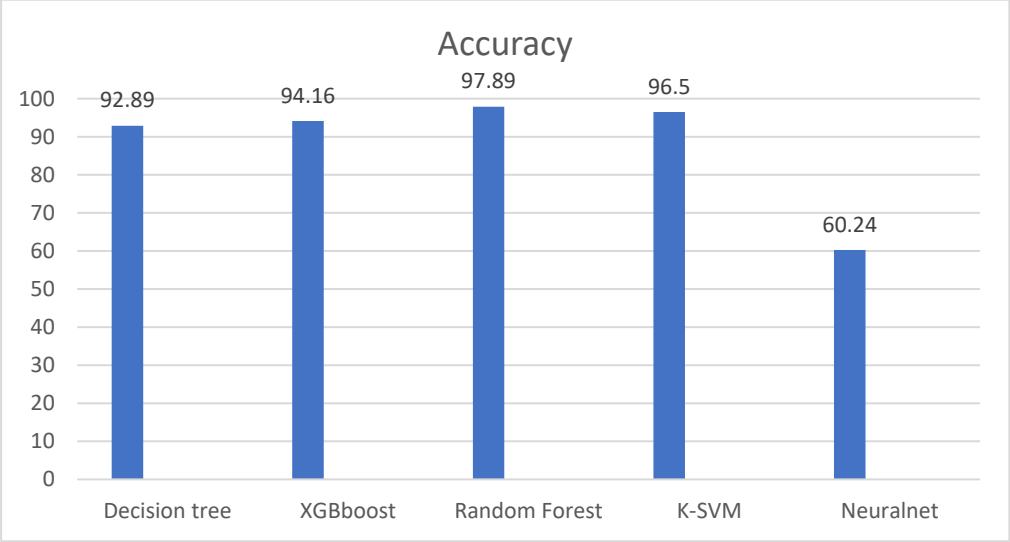

Figure 5: Comparison analysis of accuracy values of classification models

5. Conclusion

Breast cancer is the deadliest disease rising over time amongst different countries. Lack of awareness of disease can be the main reason for more death rates. Practitioners sometimes may do mistakes because of less experience or poor investigation of reports. So, machine learning algorithms provides efficient way to predict cancer. We predict cancer using five machine learning models, where Random Forest performs best than others. In future, we analyses cancer diseases at early stage.

References

1. C. Shravya, K. Pravalika and Shaik Subhani, "Prediction of breast cancer using supervised machine learning techniques," International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.6, pp.1106-1110, 2019
2. S. R. Lakhani, E. I.O., S. Schnitt, P. Tan, and M. van de Vijver, WHO classification of tumours of the breast, 4th ed. Lyon: WHO Press, 2012.
3. P. Boyle and B. Levin, Eds., World Cancer Report 2008. Lyon: IARC, 2008.
4. B. Stenkvist, S. Westman-Naeser, J. Holmquist, B. Nordin, E. Bengtsson, J. Vegelius, O. Eriksson, and C. H. Fox, "Computerized nuclear morphometry as an objective method for characterizing human cancer cell populations," Cancer Research, vol. 38, no. 12, pp. 4688–4697, 1978.
5. D. Hanahan and R.A. Weinberg, "Hallmarks of cancer: the next generation," vol.144, pp 646–74

Copyright © 2019 Author

ISSN: 0971-1260
Vol-22-Issue-17-September-2019

6. K. Kourou, P. Themis, K. Exarchos, P. Konstantinos, P. Exarchos, V. Michalis, V. Karamouzis, and F.I. Dimitrios, "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, vol. 13, pp.8-17, 2015.
7. J. R. Quinlan, "Induction of decision trees," Mach Learn, vol.1, no.1, pp. 81-106, 1986.
8. S. Kanta and A.N, Sarkar, "Identifying patients at risk of breast cancer through decision trees", *International Journal of Advanced Research in Computer Science*, vol. 08, pp. 88-96, 2017
9. L. Breiman, "Random forests," Mach Learn, vol. 45, pp. 5–32, 2001.
10. A. Hapfelmeier and K. Ulm, "A new variable selection approach using random forests," Computational Statistics & Data Analysis, vol.60, pp.50-69, 2013.
11. P. Zhang, "Neural networks for classification: A survey," IEEE Trans. Syst., Man, Cybern., Applications and Reviews, vol. 30, no. 4, pp.451-462, 2000.
12. J.A. Cruz, and D.S. Wishart, "Applications of machine learning in cancer prediction and prognosis," Cancer Inform, vol. 2. pp. 59–77, 2006
13. A.T. Azar, and S. A. El-Said, "Performance analysis of support vector machines classifiers in breast cancer mammography recognition," Neural Computing and Applications, vol. 24, no.5, pp. 1163–1177, 2014.
14. M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks, Information Processing and Management, vol.45, no. 4, pp. 427–437, 2009.

Copyright © 2019 Author