Machine Learning Techniques for Detecting and Predicting Breast Cancer

Rati Shukla, Vikash Yadav, Parashu Ram Pal, Pankaj Pathak

Abstract: ***Breast cancer is a syndrome that causes hues numbers of casualty every year due to ineffectiveness of proper filtering and appropriate classification methods. Breast Cancer is not one of the homogeneous diseases that differ greatly among different categories of Cancer sufferer and even within each individual tumor. Classification of cancer sufferer using Machine Learning methodologies in different class of risk criterion such as high, low and medium has led many research dimensions of life science data. Therefore, Machine Learning is one of the very use full methodologies to study and design the different class of development and prognosis of cancerous situation. Machine learning methods are very powerful and effective tool for key feature extraction and classification form complex cancerous data set. In this study, we put forward applicability of different Machine Learning classification techniques employed in the prediction and prognosis of Breast Cancer.***

Index Terms: Breast Cancer, Classification, Neural Network, Support Vector Machine, Cancer Susceptibility

I. INTRODUCTION

Breast cancer thrives with the breast cell. The first traces of breast cancer are liposuction and abnormal mammogram. An early warning signs of breast cancer any change in nipple size, spoon, nipple removal, and abnormal discharge of blood, temperature. The rate of breast cancer is much higher in developed country than developing country. The researcher assumes that lifestyle (unbalanced diet, physical activity) and mental problems like, excessive stress, sadness affect quality of life of cancer sufferer. To select effective treatments for chronic disease such as breast cancer patients, it is essential to carefully show on the risk and benefits of each treatment [1]. Machine learning approaches are one of the efficient ways to categorize the cancerous patient data on the basis of different symptoms. Many researchers have been conducted research to carry out machine learning approaches on various Biological datasets for cataloging [2][3].

Breast cancer is became one of the major reason of death among females [4]. For identifying breast cancer symptoms mammography is one of the effective techniques but the quality of mammography technique is very poor. Hence, it is not easy to interpret mammography image. The important abnormalities in breast cancer are Calcification and masses. Micro calcification defined as the high frequency part and noise having a lower frequencybackground.

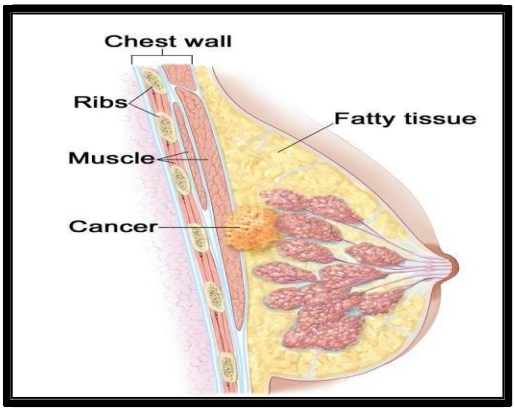

Fig. 1:Representing Cancerous Tissues and Fatty Tissues

The challenges in correct segmentation of each micro calcification are because of the variation in shape and size of micro calcification and presence of high frequency noise in superimposed surrounding tissue [4]. The Prognosis is an important point that mostly physician and cancer sufferer communicate as complicated to argue and one of the ways of suggests prognostic detail to improve sufferer understanding [6] [7]. Pathological approaches for determining lumps status and their symptoms are useful for breast cancer prediction and prognosis. Machine learning prediction techniques have benefited health care industry a lot. It suggests different class of change and watch tool, targeted at improving cancer sufferer' protection and wellness program [8]. Medical databases are too complex to handle. There are several investigations on the medical database. Pre-processing is necessary for real-time medical data. Improve medical data with a big trait, with an effective feature selection algorithm to calculate binary and multiply data [9].

The main goals of array express are repository for data, high quality gene expression and experimental protocols [11]. One of the important dimensions of bioinformatics research is analysis of coherent patterns in gene expression data. [12][13]. Microarray techniques are very valuable and powerful, so it is important to extract the greatest value from the microarray data, especially from the larger microarray sample series [14]. Many branches, like decision-making, financial and medical research can use Multi-relational classification for the destination [16] [17] [19] [26].

Revised Manuscript Received on May 06, 2019

Rati Shukla,GIS Cell, MNNIT Allahabad
Vikash Yadav, Department of Computer Science and Engineering, ABES Engineering College, Ghaziabad
Parashu Ram Pal, Department of Information Technology and Engineering, ABES Engineering College, Ghaziabad.
Pankaj Pathak, Symbiosis International (Deemed University), Pune, Maharashtra

II. Problem and Challenges

To construct efficient and right classifiers for biological application in machine learning areas is one of the leading challenges [4]. There are lots of requirements of computational biologists to handle and assist translate the big amount of data that is steadily being accumulated in the genomic study [10]. Genes and protein expressions, ways to analyze high-transfer data in the form of photos and images are computing techniques becoming important for understanding the diseases and importance of discovering the drug in the future [15]. Classification of genes expression using Machine learning is a research area that presents a new challenge because of the unique peculiarity of the problem [18]. There are mainly following challenges in genes expression classification using machine learning approach, the hues number of gene expression, relevant features investigation, and occurrence of noises inherent in the dataset, classification accuracy and reliability

III. Classifier for Breast Cancer Prediction

Machine learning makes the use of data mining algorithms to find patterns in large datasets. Number of breast cancer prediction models has been developed using statistical and machine learning techniques and employed. Artificial intelligence has taken a great place in the scientific and technical development community [21]. Effective uses of Machine learning classification based data extraction techniques are available like Random Forecast (RF), Support Vector Machines (SVM), Naive Bayes Classifier (NBC), Decision Trees (DT), K-Nearest Neighbor (KNN), Logistic Regression, Artificial Neural Networks (ANN) to massive volume of healthcare data. ML methods, is generally used to show covert correlations between diseases and gene expression. Timely detection and right investigation of the problem using ML technique will help the doctor in saving the life of a cancer patient. Due to varied appearances and complexity of lumps, the Brain Tumor Detection based on Machine Learning algorithms method gives the satisfied accuracy. It requires high degree of accuracy as human life involved. Number of ML techniques has been applied to widely find characteristic of disease diagnosis and prediction. By learning machine, physician can check the clinical performance of the drug against cancer by transferring the facilities obtained from the data on the basis of cell lines for personal patients. A classification model built based on gene expression measurements of samples from patients who have cancer on the left, right, and both lobes of the prostate as classes. Classify different cancerous cases using standard machine learning strategy to find genes are valuable and effective way to analysis

A. Naive Bayes Classifier (NBC)

NBC is one of the efficient Machine learning Algorithm based on Bayes theorem with independent postulation with predictors for classification problems. Probabilistic Approach to Classification is relationship between input features and class expressed as probabilities.

$$P(A/B) = \mathrm{P}\frac{\left(\frac{\mathrm{B}}{\mathrm{A}}\right)\mathrm{P(A)}}{\mathrm{P\,(B)}}$$

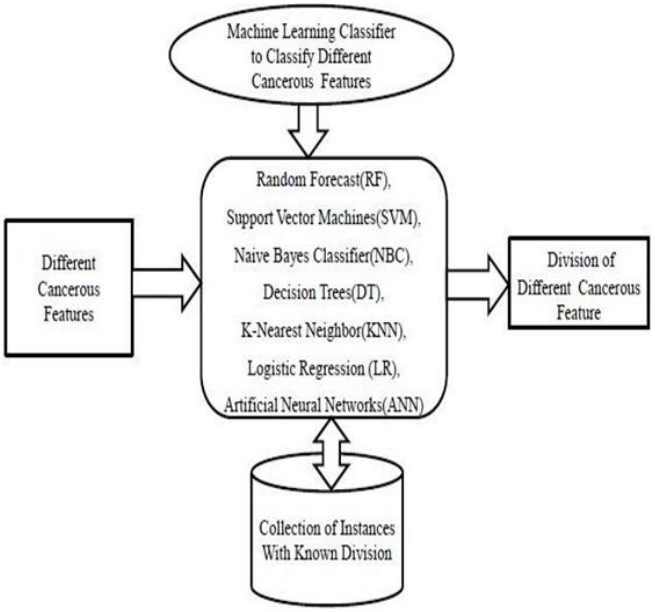

Fig. 2:Machine Learning Classifier to Classify Cancerous Features

P (A/B): The conditional probability of occurrence of event A given the event B is true. P (A) and P (B): The Probability of Occurrence of event A and B respectively. P (B/a): Conditional Practicability of occurrence of event B given the event A is true.

B. Logistic Regression (Predictive Learning Model)

A type of supervised machine learning approach for prediction model. Logistic Regression Generalize idea of integer regression to situations where outcome variable is categorical. This technique mostly focuses on binary classification of data.

$$P(Y = 1) = \alpha_0 + \alpha_1 X$$

$$P(Y=1) = \frac{e^{\beta_0} + e^{\beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

P(Y=1): Predicted probability above 1 or below zero. Logistic Regression is a generalized idea of linear regression to situation where outcome variable are categorical. Logistic Regression is a predictive Learning Model when the output label predicted a value between 0 andone.

C. Support Vector Machines (SVM)

SVM is now days one of the most effective Machine learning classifier under the supervised machine learning techniques used for classification or regression challenges in the field of medical sciences. Supports Vector Machines are based on two key concepts are selection of hyper plane which segregate the two classes and maximum distance between the nearest data points of margin. A hyper plane is one that separates between a set

$$\sum_{k} (\quad)^2$$

of data points having different margin memberships. A schematic example is shown in Figure-02. The objects belong either to class BLUE or BLACK [22][23].

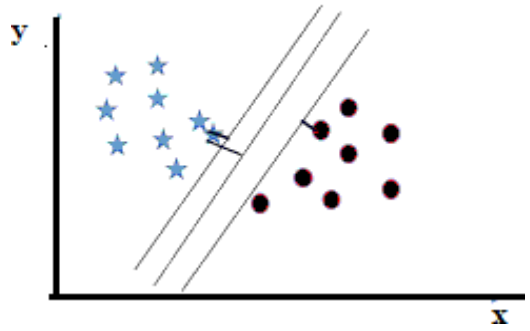

Fig 3:Support Vector Machines (SVM)

The main behind SVM based classification is to create a choice boundary between two data points of margin that enables the forecasting of labels. SVM is very supportive in managing classification tasks for high-dimensional and sparse microarray data.

D. Support Vector Machines (SVM)

DT recursively divide cancer sufferer data into different classes based on the value. Decision tree (DT) provides efficient tool for classification and forecasting in the area of life sciences [24] [25]. Assess a machine learning perspective to evaluate the activities of the abnormal thruways in lumps, which can help to identify hidden respondents.

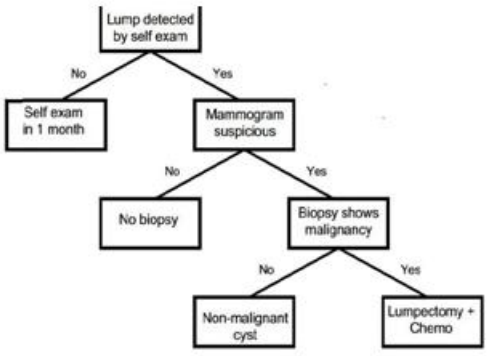

Fig 4: Decision tree for breast cancer diagnosis

E. Artificial Neural Networks (ANN)

ANN is a system inspired computational model based on the function and structure of biological nerve system. The idea behind ANN is basically set of interconnected neurons. Three main component of neural network are basically used for transformation are layer Input, output, hidden layer. The neurons are connected by means of edges and each edges is associated with their vertex (input, output, hidden) called wait.

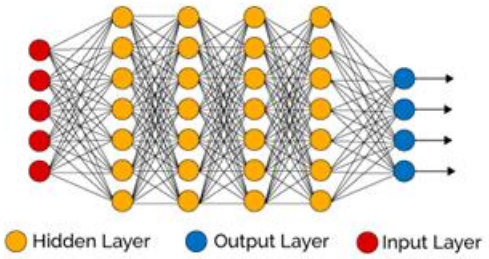

Fig 5: Deep Learning Neural Network

F. K-Nearest Neighbor (KNN)

One of the simplest algorithms that store all existing cases and classifies the new cases based on their similarity measure. KNN is A simplest non parametric method for classification based on finding the k nearest in some reference set and taking a majority poll among the margin of this. 'Nearest' can be measured by Euclidean distance method (EDM).

Euclidean Distance Function=

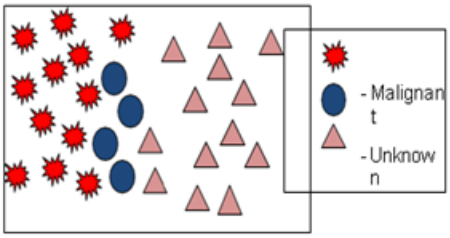

Fig 6: Cancerous Tumor Classification Using KNN Algorithm

IV. Prediction of Breast Cancer Using ML Techniques

Breast cancer can reappear month or a year after the treatment at the same place called recurrent breast cancer. Breast cancer that has returned and it spread out to other parts of the cancer suffers human body like liver, brain, lungs is metastasis or distant recurrence. Common symptoms of metastatic breast cancer sufferer May angry, feel scared, stressed, outraged, and depressed. Local recurrence is usually found on a diagnostic mammogram, during a physical examination by a cancer expert cancer sufferer see a change in breast in X-ray examination. The surgeon eliminates the lumps, diagnosed by the pathologist and tested for hormone receptor status. If physical examination symptoms are abnormal the pathological results can sure/unsure metastasis. Discomfort, nipple discharge, inverted nipple, growing view, new shape or size, nipple crust lumps, Shortness of breath, Weight loss, and Bone pains are the common symptoms. The machine based classification and feature selection methods applied before building modules for prediction of cancer recurrence.

Efficient Feature choice can directly reduce the number of original features by selecting a subset of them that still retains complete information for classification. Joint effort of different recent machine learning techniques and note of cancer surgeon/scientist for Breast cancer recurrence major future trends of computational biology seem to get good results with accuracy.

V. DISEASES PROGNOSIS

The achievement of a disease diagnosis is totally relying on quality of a non-decline medical diagnosis. Prognostic prognosis is more than that simple diagnostic investigation. Analysis of prognostic accuracy of our machine learning based approach to that of established clinical predictors and visual assessments. Cancer prognosis is mainly concerned with three predictive tasks susceptibility; recurrence survival. Among the various Machines learning Algorithm SVM provides more accurate result. Gene mutation of cancer patient's profiles effectively used with unsupervised ML methods to find clinically perceptible a division breast cancer patient's group. A good study of detecting the melanoma skin cancer using high level features of skin lesson can be taken from as well. [27]

VI. CONCLUSION

To be precise, this study addresses the importance and effective use of machine learning techniques to analyze different class of Breast cancer species and reduce the mortality rate. To analyze Biological data related to breast cancer using various machine learning, learned techniques are available. ML is actively involved in Breast cancer related complication. Good Eating habits and lifestyle influence on Breast cancer related risks. Usability of machine learning classifiers and its utility in cancer prediction/prognosis can decrease mortality. The diversified analysis of the studies focuses on development of efficient and right predictive models using supervised machine learning based classification algorithms. Application of different Machine learning classification techniques like ANN, NBC, DT, CNN, SVM techniques for feature choice and study of multiple- dimensional biological data, non-multiple data integration is a good resource for human understanding in breast cancer predictions and prognosis ofdiseases.

REFERENCES

1. Simes, R. J. (1985). Treatment selection for cancer patients: application of statistical decision theory to the treatment of advanced ovarian cancer. Journal of chronic diseases, 38(2), 171- 186.
2. Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83, 1064-1069.
3. Maclin, P. S., Dempsey, J., Brooks, J., & Rand, J. (1991). Using neural networks to diagnose cancer. Journal of Medical Systems, 15(1), 11-19.
4. Malvia, S., Bagadi, S. A., Dubey, U. S., &Saxena, S. (2017). Epidemiology of breast cancer in Indian women. Asia Pacific Journal of Clinical Oncology.
5. Gangnon, R. E., Stout, N. K., Alagoz, O., Hampton, J. M., Sprague, B. L., & Trentham-Dietz,
6. (2018). Contribution of breast cancer to overall mortality for US women.Medical Decision Making, 38(1_suppl), 24S-31S.
7. Mccarthy, J. F., Marx, K. A., Hoffman, P. E., Gee, A. G., O'neil, P. H. I. L. I. P., Ujwal, M. L., & Hotchkiss, J. (2004). Applications of Machine Learning and High Dimensional Visualization in Cancer Detection, Diagnosis, and Management. Annals of the New York Academy of Sciences, 1020(1), 239-262.
8. Hagerty, R. G., Butow, P. N., Ellis, P. M., Dimitry, S., & Tattersall, M. H. N. (2005). Communicating prognosis in cancer care: a systematic review of the literature. Annals of Oncology, 16(7), 1005-1053.
9. Nithya, B., &Ilango, V. (2017, June). Predictive analytics in health care using machine learning tools and techniques. In Intelligent Computing and Control Systems (ICICCS), 2017 International Conference on (pp. 492-499). IEEE.
10. Vanaja, S., & Kumar, K. R. (2014). Analysis of feature selection algorithms on classification: a survey. International Journal of Computer Applications, 96(17).
11. Cohen, J. (2004). Bioinformatics—an introduction for computer scientists. ACM Computing Surveys (CSUR), 36(2), 122-158.
12. Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., &Lilja, P. (2005). Array Express—a public repository for microarray gene expression data at the EBI. Nucleic acids research, 33(suppl_1), D553-D555.
13. Jiang, D., Pei, J., & Zhang, A. (2005). An interactive approach to mining gene expression data. IEEE Transactions on knowledge and Data Engineering, 17(10), 1363-1378.
14. Korenberg, M. J. (Ed.). (2007). Microarray data analysis: methods and applications (Vol. 377). Springer Science & Business Media.
15. Zhang, Y., &Rajapakse, J. C. (2009). Machine learning in bioinformatics (Vol. 4). John Wiley & Sons.
16. Han, J., Pei, J., &Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.
17. Yin, X., Han, J., Yang, J., & Philip, S. Y. (2006), Crossmine: Efficient classification across multiple database relations. In Constraint-Based mining and inductive databases (pp. 172- 195). Springer, Berlin, Heidelberg.
18. Yin, X., Han, J., Yang, J., & Yu, P. S. (2006). Efficient classification across multiple database relations: A crossmine approach. IEEE Transactions on Knowledge and Data Engineering, 18(6), 770-783.
19. Lu, Y., & Han, J. (2003). Cancer classification using gene expression data. Information Systems, 28(4), 243-268.
20. Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., &Saez- Rodriguez, J. (2013). Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. PLoS one, 8(4), e61318.
21. Agrawal, R., &Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).
22. Kuo, C. Y., Yu, L. C., Chen, H. C., & Chan, C. L. (2018). Comparison of models for the prediction of medical costs of spinal fusion in Taiwan Diagnosis-Related Groups by machine learning algorithms. Healthcare informatics research, 24(1), 29-37.
23. Jiang, Y., Xie, J., Han, Z., Liu, W., Xi, S., Huang, L., & Yu, J. (2018). Immuno marker Support Vector Machine Classifier for Prediction of Gastric Cancer Survival and Adjuvant Chemotherapeutic Benefit. Clinical Cancer Research, clincanres-0848.
24. Vapnik, V. (2013). The nature of statistical learning theory. Springer science & business media.
25. Elsayad, A. M., &Elsalamony, H. A. (2013). Diagnosis of breast cancer using decision tree models and SVM. International Journal of Computer Applications, 83(5).
26. Elsalamony, H. A., &Elsayad, A. M. (2013). Bank Direct Marketing Based on Neural Network and C5. 0 Models. International Journal of Engineering and Advanced Technology (IJEAT), 2(6).
27. Yadav, V. & Kaushik, D. V., (2018). A Study on Automatic Early Detection of Skin Cancer. International Journal of Advanced Intelligence Paradigms (IJAIP), ISSN online: 1755-0394 ISSN print: 1755-0386, Vol. 12, No. 3/4, pp. 392-399, March 2019, U.K., DOI: 10.1504/IJAIP.2018.10015438.
28. Yadav, V. & Kaushik, D. V., (2018). Detection of Melanoma Skin Disease by Extracting High Level Features for Skin Lesions. International Journal of Advanced Intelligence Paradigms (IJAIP), ISSN online: 1755-0394 ISSN print: 1755- 0386, Vol. 11, Nos. 3/4, pp. 397-408, September 2018, U.K., DOI: 10.1504/IJAIP.2018.10015438.

AUTHORS PROFILE

Rati Shukla received her BSc in 2006, MCA degree in 2009 from U.P. Technical University Lucknow and MTech (Computer Science and Engineering) degree in 2014 from Motilal Nehru National Institute of Technology, Allahabad (U.P. India). She is pursuing her PhD at GIS Cell Motilal Nehru National Institute of Technology, Allahabad (U.P. India). She worked as a Guest Faculty at the Department of Computer Science and Engineering, Motilal Nehru National Institute of Technology, Allahabad (U.P. India) from 2010 to 2012. Her areas of interest are genetic algorithm, data structure.

Dr. Vikash Yadav received his B.Tech (Computer Science & Engineering) degree in 2009 from Dr. Ambedkar Institute of Technology for Handicapped, Kanpur (U.P. India), M.Tech (Software Engineering) degree in 2013 from Motilal Nehru National Institute of Technology, Allahabad (U.P. India) and Ph.D (Computer Science & Engineering) degree from Dr. A.P.J Abdul Kalam University (Formerly U. P. Technical University) Lucknow, (U.P. India) in 2017 in the field of Image Processing. He is currently working as an Assistant Professor in the Department of Computer Science & Engineering, ABES Engineering College, Ghaziabad, India and has more the 7 years of Teaching/Research experience and published more than 30 research papers in various National/International Conferences/Journals. He is also a reviewer of various SCI/SCIE/Scopus indexed journals. His area of interest includes Data Structure, Data Mining, Image Processing and Big Data Analytics.

Dr. Parashu Ram Pal, obtained Masters and Ph.D. in 1998 and 2010 respectively. He is working as a Professor in Department of Information Technology, ABES Engineering College, Ghaziabad, India. His area of interests are DBMS, Data Mining, Automata Theory, Computer Graphics and Computer Architecture. He has published more than 30 Research Papers in various International, National Journals & Conferences. He is devoted to Education, Research & Development for more than twenty years and always try to create a proper environment for imparting quality education with the spirit of service to the humanity. He believes in motivating the staff and students to achieve excellence in the field of education and research.

Dr. Pankaj Pathak obtained Masters and Ph.D. in 2005, 2014 respectively. He is working as an Assistant Professor in Symbiosis Institute of Telecom Management. His area of interests are Data Mining, AI, and Smart Technologies. He has Published Several Research papers in the area of Data Mining, IOT security and Speech Recognition Technology.