Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 4): pp. 3933-3941
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.04.430

Breast Cancer Prognosis And Detection: A Comparative Study Of Supervised Machine Learning Approaches

Neelam Singh, Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, India neelamjain.jain@gmail.com

Vijay Laxmi Thapliyal, Department of Computer Application, Graphic Era Deemed to be University, Dehradun, India laxmithapliyal100@gmail.com

Vandana Rawat, Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, India vandanarawat@geu.ac.in

Umang Garg, Associate Professor, Department of Computer Science and Engineering, Graphic Era Hill University, Dehradun.

Abstract: Cancer is one of the most dreadful disease taking heavy toll of human life in spite of advances in the field of medical science. Among all type of cancer, Breast Cancer is amongst the most usual category affecting women everywhere in the world and it is amid the foremost reason of death toll in women. A careful selection of techniques are required to analyse data and generate accurate results. Efficient techniques and methods are required to analyse data for accurate decision making and prediction. Machine Learning algorithms has achieved a bench mark when examination of data set is concerned for predictive analytics. Researchers and scientific community are working to achieve higher accuracy rate to predict ailments like breast cancer. Every technique and algorithm provide varying accuracy for different data sets and tools. In this study we will do a comparative investigation of different algorithms to find the highly appropriate and accurate breast cancer prediction algorithm. Algorithms like KNN, Decision Tree, SVM, Random forest are being used for the study.

Keywords: Machine Learning, Predictive Analytics, KNN, Decision Tree, SVM, Random Forest.

1. Introduction

Among cancer, breast cancer is commonly found in women and it is a fatal disease that cause huge number of casualties of women all around the globe. According to (WHO) "World health organization's" part "International Agency for Research on Cancer (IARC)". The number of casualties done by cancer was 8.2 million alone in 2012. That's makes it the second largest cause which is responsible for women casualties and the cases of

breast cancer are increasing with high speed. Specialists have found some causes that can leads to breast cancer like hormonal imbalance, lifestyle, surrounding environment that may increase the chances of an individual of developing breast cancer. In some cases patients have associated to gene mutations that cause breast cancer. There are some ways that can be used to diagnose breast cancer by testing medical images using radiology and histology images. The radiology images examination can prove beneficial to recognise the regions of malformation. However, they cannot be applied to determine whether region is cancerous or not. Examine the tissue taken from patient under microscope is the only sure way to determine the regions is cancerous or not. The histology images allows us to make a distinction between normal tissue, benign and malignant lesions Generally, histopathologists inspect the consistencies of cell shapes and determine cancerous areas visually. But this may lead to a false diagnosis, In the event that the histopathologists are definitely not all around trained. Therefore there is an emphasize desire for computer-based diagnosis. For the prognosis of breast cancer there are multiple machine learning techniques that are being utilized. Applying an efficient and accurate technique for knowledge discovery is one of the prime task in prediction. It commonly occurs when affected uncommon cells started spreading other places of body organs. No such breast cancer prevention is there, but the chances of patient survivability can increases through early discovery of disease. Although Machine learning can be a good alternative way to recognise breast cancer.

In the field of medical, Data (It could be structured or unstructured data) about patients of various diseases are collected on daily basis. For example Structured data –it could be patient's gender, age, height, weight etc. and unstructured Data- it could be patient's readme illness and medical history. Processing these datasets and finding hidden patterns and valuable knowledge will improve and make the medical service and healthcare better. As computer science and algorithms is developing rapidly that has permitted for innovative methodologies to use data in order to discover more insight to take advantages in competitive world, One of the fastest growing domain of computer science is machine learning. Its foremost concern is to enable computers to learn from its experience using input or also known as training data, and extract knowledge from input data in order to perform various tasks. . Machine learning includes supervised, unsupervised and reinforcement learning. Machine learning methods which are used with the data samples are describes in context of attribute or in other word features, which might be of various values and types. Analysis of enormous data sets is tough when it points to acquire more correct and dominant patterns which can lead us to a result and information that enable process automation, decision making and enhanced insight.

2. Literature Survey

Dalal Barou et al. [1] presented a CNN–based multimodal for disease prediction algorithms where they collected real time data from central china hospital in 2013-2015 for that purpose they used both structured and unstructured data. To deal with

structured data useful features are extracted with the advice of domain experts and for unstructured data they select features automatically and they achieved accuracy 94.8%.

Noreen Fatima et al. [2] proposed an analysis through comparison of various algorithm of machine learning where they comparatively analysed several kind of machine learning (supervised and unsupervised), deep learning and data mining methods to search the utmost fitting techniques which can uphold the enormous set of data with correctness of prediction.

Anji reddy vaka et al. [12] projected a novel method for detecting breast cancer called DNNS. This method is presented to produces better quality images and other parameter which is related to performance. In experimental result proposed DNNS has proved that it is quite better than the existing methods.

Hasib iqbal et al. [13] presented an analysis through the comparison of various algorithms for breast cancer prediction of machine learning and five algorithms are SVM, KNN, ANNs, Random forest and logistic regression and ANNs obtained the highest accuracy that is 98.57%.

Chaurasia et al. [17] implemented data mining algorithms specifically NB, RBF network, and J48 DT on large dataset. WEKA version 3.6.9 tool was used for analysis in their research. The results shows that 97.36% of accuracy obtained by Naïve bayes which is more than RBF network and J48 DT that is 96.77% and 93.41%, respectively.

Banu et al. [18] presented Bayes classifiers performance like through comparison study. For implementation of the models they used Statistical Analytical Software Enterprise Miner (SAS-EM). Where BAN, BBN and TAN achieved 91.7%, 91.7%, and 94.11% accuracy respectively. Among Naïve Bayes techniques TAN is the best classifier for this data suggested by their research.

Yeu et al. [19] mainly presented reviews on different machine learning algorithms and Wisconsin Breast Cancer Diagnosis dataset was drawn for that purpose. According to them, ANN architecture and deep belief networks approach has obtained 96.68% accuracy, 99.10% of classification accuracy has achieved by SVM technique. Ensemble technique was also reviewed and using voting technique some algorithms were implemented. The ensemble technique obtained 93.13% accuracy.

3. Algorithms

3.1. LOGISTICS REGRESSION (LR)

Logistic regression is applied to analyse a dataset in which a result is decided by single or more independent variables. The result can only be measured with a binary variable (only two result are possible) that is either 1 or 0.

3.2. KNN

KNN is a supervised regression and classification machine learning algorithm [15]. It means for regression as well as classification predictive problem it can be used and K nearest neighbour is non-parametric method and works on a simple approach it categorize a data point depend on the point that is most common to it. For that purpose it uses Euclidean distance.

3.3. DECISION TREE (DT)

It is based on regression and classification model, based on certain conditions it provides all the possible solutions through graphically [14]. "A decision tree is similar to tree structure, where node shows a test on an attribute, Branch signify an outcome of the test, and a class label is represented by each leaf node" [14]. Dataset is further divides into small number of branches and each branch represent an outcome which is used for prediction.

3.4. NAIVE BAYES (NB)

This algorithm is based on bayes theorem [14]. It follows some assumption like every pair of feature are independent and every given feature is equally important. For large training dataset this model is particularly make an assumption. It is used to determine the possibility or probability with the help of Bayesian method and It is a powerful algorithm for predictive analysis. Bayes theorem is expressed with the help of given equation.

$$P(A/B) = \frac{P(B/A)P(B)}{P(A)} \qquad (1)$$

3.5. SUPPORT VECTOR MACHINE (SVM)

SVM is used for regression as well as classification problems and it is a supervised learning algorithm. The main motive behind in SVM algorithm is to find a hyperplane (basically they are decision boundaries) in N- dimensional space that classifies the data points. Hyperplanes are helpful in order to classify the data points. In two-dimensional space, data points are divided by hyperplane into two segments [14]. While doing prediction with large datasets SVM can provides the highest accuracy rate.

3.6. RANDOM FOREST (RF)

This algorithm [16] is a supervised learning method that is widely exploited to resolve regression and classification problems. In this classifier, The high accuracy results achieved through the higher number of trees. It doesn't over fits the model, if there are many trees in random forest. But however, it is generally used for classification problems. It is widely utilize for prediction of output value based on previous given input.

4. Proposed Methodology

In order to perform comparative analysis the behaviour of various algorithms is needed to compare, to perform breast cancer prediction we followed methodology shown in figure 1.

4.1 Data collection

For that purpose, UCI machine learning repository dataset was retrieved for breast cancer comparative analysis. There are total 569 instances of this dataset including 357(63%) are diagnosed as the benign type and 212(37%) are diagnosed as the malignant type.

4.2 Data-pre-processing

Our dataset might be incomplete or have few missing values, which can manipulate our accuracy if it is not processed, in data pre-processing we process raw data and transform it in order to handle it in an effective way.

The complete process is shown in the flow diagram given in Figure 1.

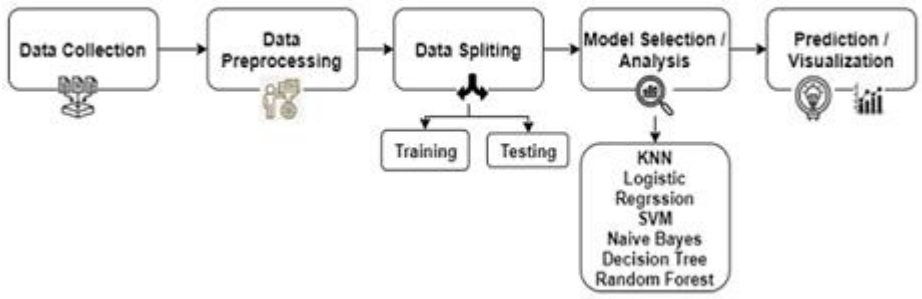

Figure 1 Proposed Methodology flow diagram

4.3 Data splitting

In data splitting, dataset is distributed into two subsets basically referred to as training data & testing data.

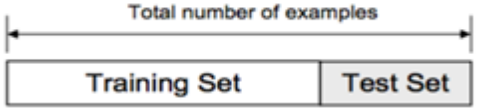

Table 1 and figure1 shows the accuracy of different algorithms when dataset is divided i nto 75% & 25%

Table 1 Accuracy after 75% and 25% data split

Algorithm	Accuracy

LR	95.10%
KNN	95.10%
SVM	96.50%
NB	94.4%
DT	93.70%
RF	96.50%

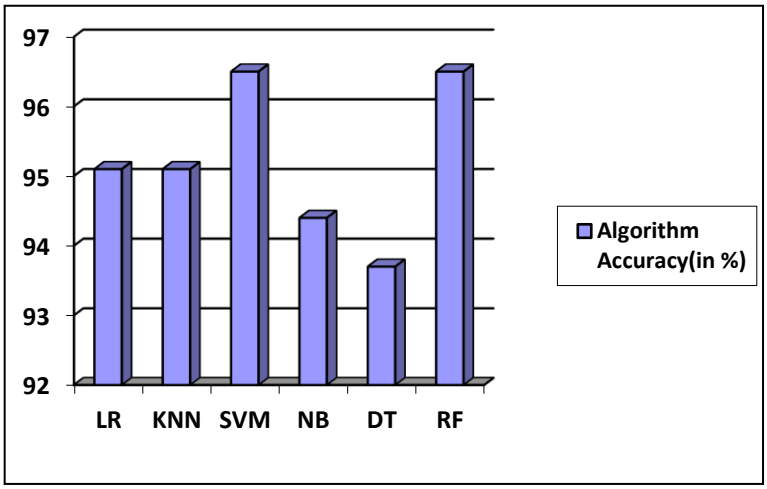

Figure 1. Accuracy after 75% and 25% of division

Table 2 and figure 2 illustrates the accuracy of various algorithms when dataset is divide d into 80% & 20%

Table 2 Accuracy after 80% and 20% data split

Algorithm	**Accuracy**
LR	96.49%
KNN	96.49%
SVM	98.24%
NB	93.85%
DT	93.85%
RF	97.39%

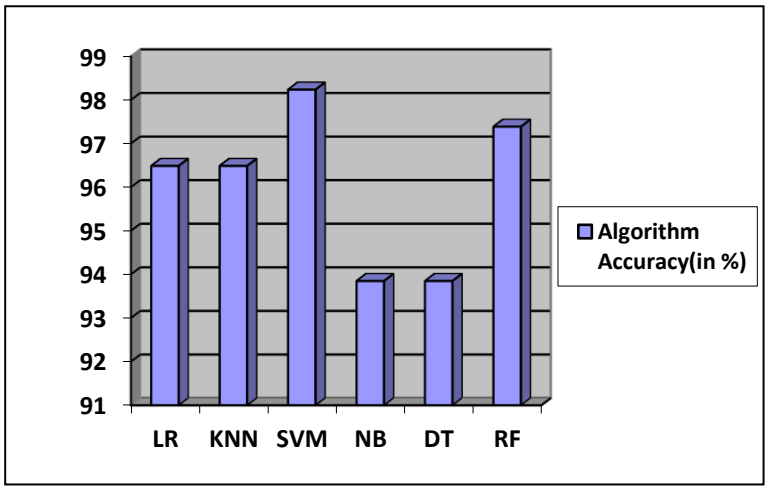

Figure 2 Accuracy after 80% and 20% of division

5. Result and discussion

For accurate prediction of breast cancer we focused on data splitting phase, firstly we divided dataset into 75% of training data and 25% of testing data all algorithms (figure 1) performed well and accuracy of two algorithms was same that is 96.50%.After that dataset was divided into 80% of training data and 20% of testing data. Accuracy of many algorithms (figure 2) was increased by 2% before it was 96.50% and now it is 98.24% and in some algorithms like logistic regression, k-nearest neighbor & random forest classifier was increased by 1% and decision tree has a slight change and naive bayes is decreased by 1%.

6. Conclusion

In this study various supervised machine learning algorithms for breast cancer prognosis and detection by focusing on the data splitting phase to attain maximum accuracy. These algorithms have played a significant role in machine learning and have been used for various purpose in different domains. In medical domain they can be very helpful by providing an early prediction of a disease and can save a life. So, now the question is which algorithm provides the highest accuracy? To address this question we have done a comparative analysis of some of the major algorithms like Naïve Bayes, Logistic Regression, KNN, SVM, Decision Tree and Random Forest by considering data splitting as one of the important parameter. We divided our dataset firstly into 75% of training and 25% of testing data and in the second phase we did the prediction by splitting the dataset into 80% of training and 20% of testing data. These changes had shown a remarkable changes in the result accuracy. Some of the algorithm performed well but Support vector

machine proved its capability in this prediction and achieved the highest accuracy 98.24% among all. This study mainly focused on the data splitting but other phases and parameter can also be further investigated to enhance the accuracy and efficiency of the prediction.

Acknowledgements: This research received no external funding.

Authors Contribution: Each author has participated and contributed sufficiently to take public responsibility for appropriate portions of the content.

Conflicts of Interest: The authors declared no conflicts of interest.

References

[1] Bardou, D., Zhang, K., & Ahmad, S. M. (2018). Classification of breast cancer based on histology images using convolutional neural networks. IEEE Access, 6, 24680-24693.

[2] Alghunaim, S., & Al-Baity, H. H. (2019). On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context. IEEE Access, 7, 91535-91546.

[3] Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. IEEE Access, 8, 150360-150376.

[4] Waseem, M. H., Nadeem, M. S. A., Abbas, A., Shaheen, A., Aziz, W., Anjum, A., ... & Shim, S. O. (2019). On the Feature Selection Methods and Reject Option Classifiers for Robust Cancer Prediction. IEEE Access, 7, 141072-141082.

[5] Aibe, N., Karasawa, K., Aoki, M., Akahane, K., Ogawa, Y., Ogo, E., ... & Sekine, H. (2019). Results of a nationwide survey on Japanese clinical practice in breast-conserving radiotherapy for breast cancer. Journal of radiation research, 60(1), 142-149.

[6] Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. IEEE access, 2, 652-687.

[7] Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. Ieee Access, 5, 8869-8879.

[8] Sivakami, K., & Saraswathi, N. (2015). Mining big data: breast cancer prediction using DT-SVM hybrid model. International Journal of Scientific Engineering and Applied Science (IJSEAS), 1(5), 418-429.

[9] 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.

[10] Dutta, S., & Bandyopadhyay, S. K. (2020). Early Breast Cancer Prediction using Artificial Intelligence Methods. Journal of Engineering Research and Reports, 48-54.

[11] Battineni, G., Chintalapudi, N., & Amenta, F. (2020). Performance analysis of different machine learning algorithms in breast cancer predictions. EAI Endorsed Transactions on Pervasive Health and Technology, 6(23).

[12] Reddy, A., Soni, B., & Reddy, S. (2020). Breast cancer detection by leveraging Machine Learning. ICT Express.

[13] Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: a comparative study using machine learning techniques. SN Computer Science, 1(5), 1-14.

[14]Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., & Zhou, Z. H. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.

[15] Gasso, G. (2019). Logistic regression.

[16] Qi, Y. (2012). Random forest for bioinformatics. In Ensemble machine learning (pp. 307-323). Springer, Boston, MA.

[17]Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology, 12(2), 119-126.

[18] Bazila Banu, A., & Thirumalaikolundusubramanian, P. (2018). Comparison of Bayes classifiers for breast cancer classification. Asian Pacific journal of cancer prevention: APJCP, 19(10), 2917.

[19] Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. Designs, 2(2), 13.