HORIZONTAL ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR DETECTION OF BREAST CANCER

Jyoti Kadadevaramath
Department of Computer Science, Government First Grade College, Dharwad-580008, Karnataka, India.
*Corresponding author E-mail: jyoti.kmath@gmail.com

Abstract:

Breast Cancer is second largest cancer in women mortality. Spotting of Breast Cancer (BC) in its early stage is the first step of diagnosis. In this regard Machine Learning Algorithmsplays capital importance in prediction and detection of breast cancer. Various supervised Classificationalgorithms are used to analyze the data. This article emphasis on comparative analysis of profuse machine learning classification algorithms to find out the mosteffective with respect to confusion matrix, accuracy and prediction. The algorithm with high accuracy can be used as best model to classify breast cancer as benign or malignant. Classification algorithms namely Logistic Regression (LR), DecisionTree (DT), Random Forest (RF), K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) are used to analyze the data.

*In the past few years, gynecological cancers have taken their toll on women's health. Breast Cancer is second largest cancer in women. Breast cancer is the major cause of cancer death followed by ovarian cancer and others. Cancer is a set of diseases that imply the abnormal growth of cells in the body. It is a disease that has harmed numerous lives and will likely continue to do so. It is fatal in many situations. One of the most common diseases, breast cancer is affected by irregular breast cellproliferation that is out of control [1]. It is usual that occasionally the abnormalities are ignored or misclassified due to the characteristics of breast malformations and the nature of human visual perception. Due to the woman's false sense of security caused by the breast's lack of discomfort, clinical breast cancer identification is a challenging undertaking [2]. It is less likely.*to be cancer and more likely to be benign cysts when the breast lump is moveable. Early detection and accurate diagnosis have the capacity to result in a full recovery and avert fatalities[3]. Early diagnosis significantly improves the chances of surviving cancer. Unfortunately, pathological analysis is a difficult, time-consuming process that demands in-depth comprehension [4]. Among other techniques for detecting breast cancer, radiologists advise utilizing digital mammograms, ultrasounds, and MRIs. Mammography is a frequently used technology because to its accessibility, low cost, and improved early detection findings [5].

For the diagnosis and detection of breast cancer, medical imaging techniques have beenwidely used. However, these techniques consume large time and require trained professional radiologists. On the other hand, the use of automated classifiers could substantially improve the diagnosis process both in terms of accuracy and time by distinguishing the image patterns automatically. Therefore, Image Processing using Neural Networks plays an important role inthe detection of breast cancer [6][7].

Around the world, the kidney stone illness has become one of the important hazardoussicknesses. It is established that, a maximum number of people are influenced by the kidney failure due to hypertension, diabetes mellitus, glomerulonephritis, etc. As kidney stone breaking can be threatening, the diagnosis of the problem in the early stage is very much essential and ultrasound imaging strategy is utilized in the medicinal practices.

However, presently kidney stone segmentation in ultrasound images has been performed manually, which is being very time consuming and depends on the expertise ofthe individual operator. In view of this, A. Nithya et. al have proposed a kidney stone detection model using artificial neural network and segmentation using multi-kernel k-means clusteringalgorithm and a maximum accuracy of 99.61% has been achieved from the experimental results compared with all other methods [8].

Method

Supervised Learning

Supervised learning uses a training set to teach models to yield the desired output. Thistraining dataset includes inputs and correct outputs, which allow the model to learn over time.The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized.

Logistic Regression

Logistic regression (LR) is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either true or False, 0 or 1, Yes or No, etc. but instead of providing the exact value as 0 and 1, it gives the probabilistic values which exists between 0 and 1. Logistic Regression is like the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. In LR, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something suchas whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc LR is a significant machine learning algorithm because it can provide probabilities and classify new data using continuous and discrete datasets. It can be used to classify the observations with different types of data and can easilydetermine the most effective variables used for the classification.

Random Forest

Random Forest is a popular machine learning algorithm that belongs to the supervisedlearning technique. It can be used for both Classification and Regression problems in ML. It isbased on the concept of ensemble learning, which is a process of combining multiple classifiersto solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decisiontrees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

Decision Tree

Decision Tree is a Supervised learning technique that can be used for both classificationand Regression problems, but mostly it is preferred for solving Classification problems. It is atree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leafnodes are the output of those decisions and do not contain any further branches.

Support Vector Machine

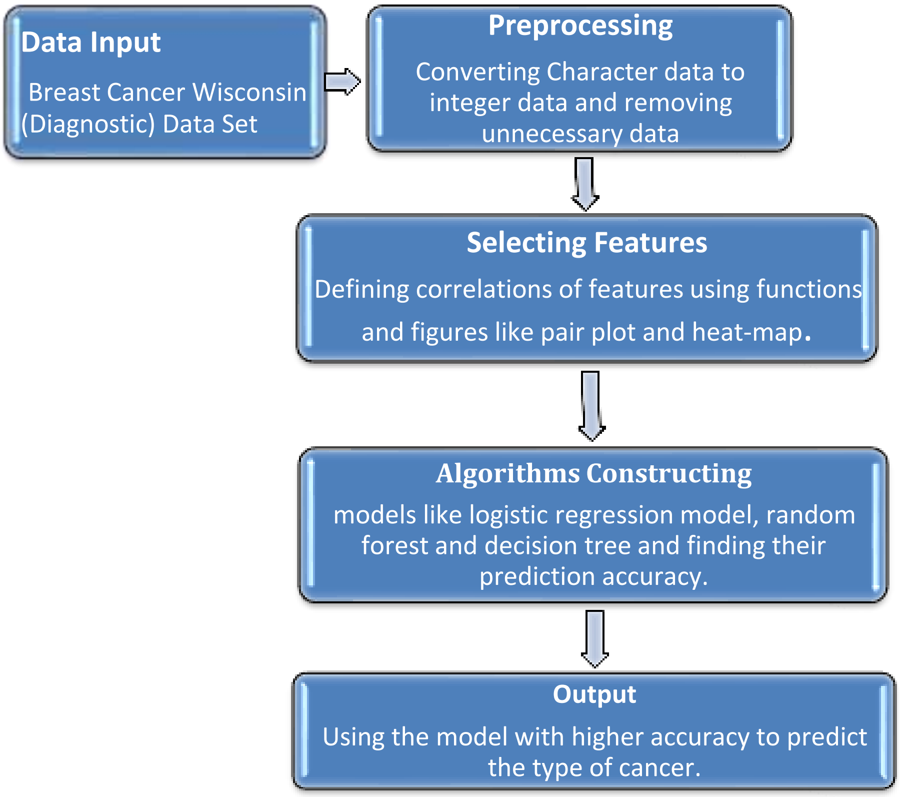
The k-nearest neighbors (KNN) algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. It is one of the popular and simplest classification and regression classifiers used in machine learning today.

K Nearest Neighbor

The k-nearest neighbors (KNN) algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. It is one of the popular and simplest classification and regression classifiers used in machine learning today.

Proposed Methodology

Work carried out on Wisconsin Diagnostic Breast Cancer (WDBC) dataset obtained from digitized images of MRI. The dataset is divided into training and testing stage for the implementation of machine learning classification algorithms. The collection includes 569 entries, 357 of which are benign (non-cancerous) and 212 of which are cancerous (malignant). In Preprocessing converting Character data to integer data and removing unnecessary data. All the work is done in the Google colab environment based on python programming language and Scikit-learn library. Flow of proposed work is show in the figure 1.



Data Input

Breast Cancer Wisconsin (Diagnostic) Data Set

Preprocessing

Converting Character data to integer data and removing unnecessary data

Selecting Features

Defining correlations of features using functions and figures like pair plot and heat-map.

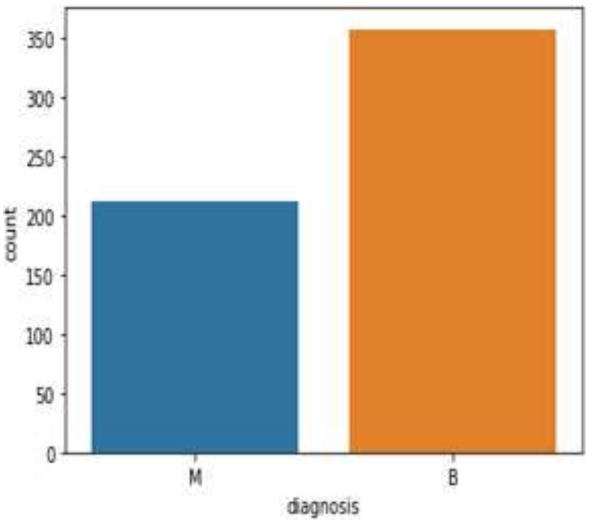
Algorithms Constructing

models like logistic regression model, random forest and decision tree and finding their prediction accuracy.

Output

Using the model with higher accuracy to predict the type of cancer.

Fig. 1: Flowchart of proposed work



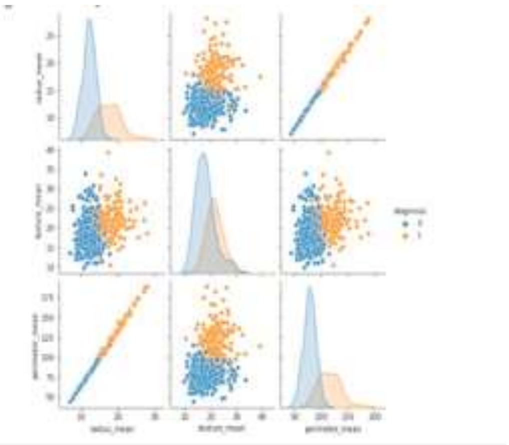
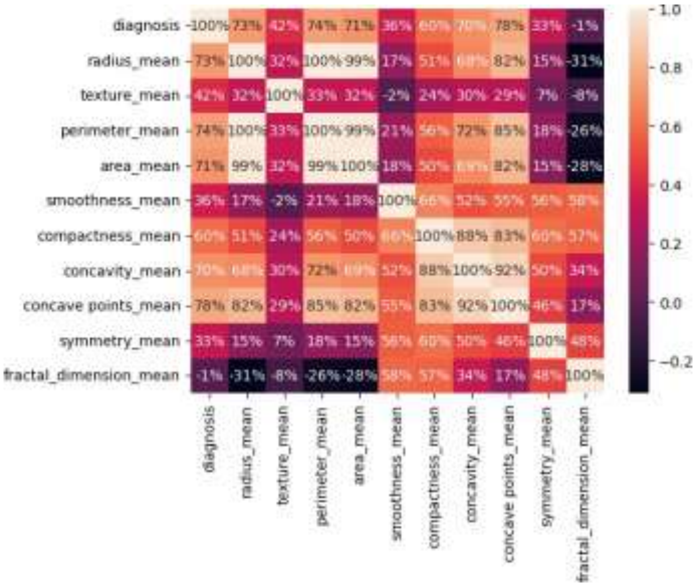


Fig. 2: Dataset Distribution

Fig. 3: Dataset Distribution

Fig. 4 :Confusion Matrix



Conclusion

This research attempted to study the comparative performances of different supervisedmachine learning algorithms in disease prediction. We only chose studies that implemented multiple machine learning methods on the same data and disease prediction for comparison. The results show that Random Forest outperformed all other classifiers with the highest accuracy.

REFERENCES

1. Heer, E., Harper, A., Escandor, N., Sung, H., McCormack, V. and Fidler- Benaoudia, M.M., 2020. Global

burden and trends in premenopausal and postmenopausal breast cancer: a population-based study. The Lancet Global Health, 8(8), pp. e1027-e1037.

2. Moloney, B.M., McAnena, P.F., Abd Elwahab, S.M., Fasoula, A., Duchesne, L., Cano, J.D.G., Glynn, C., O'Connell, A., Ennis, R., Lowery, A.J. and Kerin, M.J., 2022. Microwave imaging in breast cancer–results from the first-in-human clinicalinvestigation of the wavelia system. Academic Radiology, 29, pp. S211-S222.
3. Feng, Y., Spezia, M., Huang, S., Yuan, C., Zeng, Z., Zhang, L., Ji, X., Liu, W., Huang, B., Luo, W. and Liu, B., 2018. Breast cancer development and progression:Risk factors, cancer stem cells, signaling pathways, genomics, and molecularpathogenesis. Genes & diseases, 5(2), pp.77-106.
4. Rakhlin, A., Shvets, A., Iglovikov, V. and Kalinin, A.A., 2018. Deep convolutionalneural networks for breast cancer histology image analysis. In Image Analysis andRecognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15 (pp. 737-744). Springer International Publishing.
5. Birnbaum, J.K., Duggan, C., Anderson, B.O. and Etzioni, R., 2018. Early detectionand treatment strategies for breast cancer in low-income and upper middle-incomecountries: a modelling study. The Lancet Global Health, 6(8), pp.e885-e893.
6. M.M. Mehdy, P.Y. Ng, E.F. Shair, N.I. Md Saleh, C. Gomes, Artificial neural networks in image processing for early detection of breast cancer, Computational and Mathematical Methods in Medicine, (2017), DOI:10.1155/2017/2610628.
7. S. Bagchi, A, Huong, Signal processing techniques and computer-aided detection systems for diagnosis of breast cancer – A Review Paper, Indian J Sci. & Tech., 10(3), (2017), DOI: : 10.17485/ijst/2017/v10i3/110640.
8. A. Nithya, A. Appathurai, N. Venkatadri, D.R. Ramji, C.A. Palagan, Kidneydisease detection and segmentation using artificial neural network and multi-kernel k-means clustering for ultrasound images, Measurement, 149 (2020), 106952.