ISSN 2278-3091

Volume 9, No.5, September - October 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse334952020.pdf

https://doi.org/10.30534/ijatcse/2020/334952020

Breast cancer detection and classification approach based on ensemble learning

Bhal Chandra Ram Tripathi[1], Saksham Bhadauria [1], Krishna Prasad R [2] , Visheshwar Pratap Singh[3]

[1] Robert Bosch Engineering and Business Solutions
Bangalore, Karnataka, India, bhalchandra.chandra7@gmail.com

[2] Global Academy of Technology
Bangalore, Karnataka, India

[3] Nielsen (India) Private Limited
Mumbai, Maharashtra, India

ABSTRACT

The application of machine learning is constantly increasing especially in the field of automated disease diagnose and prediction such as breast cancer, which is very common in females as it can cause too many death but machine learning can increase the chances of survival by early prognosis and diagnosis if it diagnosed properly and accurately. In this paper we propose an automated breast cancer prediction approach based on the XgBoost random forest classifier(XGBRF) algorithm. In order to prove the effectiveness and accurateness of the proposed approach, Wisconsin diagnose breast cancer dataset is usedover which various classification rates like precision, recall , F1score and confusion matrix are generated. The testing accuracy of our proposed approach is 99%. Apart from that the proposed approach is being compared with various other approaches based on other machine learning classifiers like support vector machine, K- nearest neighbor, Naïve Bayes etc. .

Key words : Machine learning, Xgboost, Random Forest Classifier, Confusion Matrix, Support Vector Machine (SVM).

1. INTRODUCTION

Breast cancer is a type of cancer occurs in the cells of breast which can later forms a tumor. Breast cancer and the second most dangerous cancer after the lung cancer occurs commonly in female Breast cancer case is Increasing rapidly now a days as per the report provided by various institutes and organization. According to one of the record proposed by world cancer research fund(WCRF)[1] is that about 2 million cases of breast cancer are recorded in the year 2018.According to American cancer society[2] and National Cancer Institute[3] 265,000 cases in females and 2,200 cases in males are diagnosed each year which is increasing rapidly and if we will compare to deaths than 40,000 females and 440 males are died each year. In India also various cases are seen now days, but early diagnosis can reduce the chance of death due to proper treatment at correct time**.**

There are two types of breast cancer occurs i.e. one is malignant and another is benign. The application of data science and machine learning algorithm in medical fields shows that a time will come where we can diagnose various type of disease as doctors who are unable to take the decision sometimes due to unawareness but with the help of machine learning performance metrics like precision and recall are information retrieval problem through which we can extract information like how many of person have actually malignant cancer and how many person have benign as malignant cancer is more harmful then benign cancer as it spreads to all parts of our body as compare to benign cancer.

Previously several studies has been made on same research topic where they use the different machine learning algorithm for breast cancer detection such as SVM, Decision tree, Random forest, KNN classifier but in this paper an ensemble learning with XGBRF classifier algorithm on Wisconsin diagnose breast cancer dataset[5]. To increase the training and testing accuracy we have applied brute force technique and hyper parameter tuning by iteration process on each and every random state I used kfold cross validation technique too so that to get what maximum score I can get, which is elaborate in methodology itself.

2. Literature Review

This section simply encapsulates a number of correlated approaches earlier done on breast cancer diagnosis by researchers using diverse machine learning classifiers are discussed. In the year 2013, Ahmad et al. [9] come up with a paper in which the performance of decision tree (C4.5), SVM, and ANN are compared. The Iranian center for breast cancer dataset is used for the evaluation and classification. The SVM classifier delivers the highest accuracy followed by ANN and decision tree. Then in the year 2015, Nematzadeh et al. [10]presented a brief comparative analysis among the decision tree, NB, NN and SVM with three different kernel functions as classifiers for classification on the Wisconsin Breast Cancer (WBC). The evaluation result showed that SVM-RBF (10-fold) had achieved maximum accuracy of 98.32% in WPBC dataset and NN (10-fold) had achieved the highest accuracy of 98.09% over the WBC dataset. Hasan and Tahir [11], come up with an ANN classifier utilizing the PCA for the preprocessing f data as an optimal tool to enhance the differentiation in between malignant and benign tumors on WBC dataset. In this paper, three rules of thumb of PCA namely screen test, cumulative variance and Kaiser Guttman rule as feature selection are employed. Then in the year 2017, Ojha and Goel [12], come up with an approach in which they have employed various machine learning algorithms to forecast recurrent cases of breast cancer using the Wisconsin Prognostic Breast Cancer (WPBC) data set.

The SVM and decision tree (C 5.0) are the best predictors with 81% accuracy, while fuzzy c-means was found to have the lowest accuracy of 37% as per the evaluation results. Then in the year 2014, an approach was proposed for the diagnose and analysis of breast cancer disease employing two famous classifiers which are SVM and Multilayer Perceptron using Back Propagation Neural Network (MLP BPN). As per the evaluation section of this paper, the SVM was the best classifier by the Ghosh et al. [13]. In the year 2010, Osareh and Shadgar [14], come with a paper in which the issuesrelated to breast cancer diagnosis and prognostic risk evaluation of recrudescence and metastasis using SVM, K-nearest neighbor (KNN) and probabilistic neural network (PNN) are addressed. These classifiers were combined with sequential forward selection-based (SFS) feature selection, signal-to-noise ratio (SNR) feature ranking method and PCA feature transformation. The maximum overall accuracies of 98.80% was achieved by the SVM-RBF. In the year 2016, Bazazeh and Shubair [15], come up with a paper in which a comparative analysis among the SVM, random forest (RF) and Bayesian networks (BN) for breast cancer diagnosis are carried out.

The WBC dataset was used as training set to evaluate the performance of the machine learning classifiers. The evaluation section of this paper simply showcase thatthe RF had the highest probability of correctly classifying tumors while SVM had the best performance in terms of accuracy, specificity and precision. Azmi and Cob [16], come up with a system that can categorize breast cancer tumors by utilizing the neural network with feed –forward back propagation algorithm. The University of Wisconsin (UCI) dataset was used in this paper. As per the classification results, the neural network with hidden layer of 7 delivers the maximum accuracy of 96.63%. Then Gayathri and Sumathi [17], come up with a paper in which a brief comparative analysis in between the Relevance vector machine (RVM) along with others Machine learning algorithms are used for predicting the breast cancer. In order to reduce the features, linear discriminant analysis (LDA) is used whereas the data was classified by the RVM algorithm. The dataset used in this work is the WBC. The sensitivity and specificity obtained from the simulation results are 98% and 94% respectively whereas this approach delivers an accuracy of 96% which is encouraging.

3. PROPOSED WORK

3.1 Dataset used

The dataset used in training and testing phase is provided by university of wiscosin[4] in which features is calculated using computerize image which has 30 features in which 10 actual values are calculated for each cell nucleus these are radius, texture, perimeter area, smoothness, compactness, concavity, concave point, symmetry, fractal dimension[4],these features are very important for the predictions of cancer these features helps in identifying the type of cancer, we can visualize these features through different graphical technique.

3.2 Machine Learning Library

In the world of analytics we can’t perform any operation on unique identifiers and id column is the unique identifiers so id columns will be dropped from the featuresdiagnosis was my label which is classifying as benign and malignant and we will be separated features and labels from each other. Apart from this feature engineering technique is applied on these dataset to select the best features this technique named is called analysis of variance(analysis of variance) which works the same as like done by linear discriminant analysis and the package which provide this technique called as select percentile by sklearn[5] the Annova techniques can be applied on both regression and classification techniques but in this research we are doing this on classification by using Random forest classifier imported from feature selection techniques by sklearn[5]

3.3 Data visualization

Visualization[6] is a kind of communication with the data When we look at the data it is only a file in a excel sheet it does not have any interpretation with it, so we have to visualize my data in order to understand my data as it is a part of descriptive statistics where we are trying to visualize our data in visual format. It is also called exploratory data analysis in which we will do univariate, bivariate and multivariate analysis using histogram, 2d scatter plot,count plot and correlation matrix(heat map) to find the relationship between variables and to decide which model will be helpful in visualization of data.

4. ENSEMBLE LEARNING ALGORITHM

Xgboost with random forest classifier

To solve high variance issue, boosting algorithm is used where Xgboost stands for extremely gradient boosting. It is not an algorithm but it is considered as implementation of gradient boosting algorithm which uses Gradient descent as a mechanism to identify the best model. There are two types of

Xgboost classifiers provided for classification one is xgboost classifier (XGBC), another isXgboost random forest classifier(XGBRF) but we used xgboost classifier (XGBC). Over fitting occurs due to it uses decision tree algorithms which cause over fitting but we can control the over fitting by tuning the hyper parameters of Xgboost. When we used brute force technique to solve the over fitting problem, it created an another sense of the solution but as there is no generalized model available for Xgboost classifier which can solve the chance of over fitting problem rather than using k-foldcross validation or grid search technique.

To overcome this problem we can use Xgboostrandom forest classifier(XGBRF)which is applicable for giving us generalized model and by applying brute force technique through which we can get the highest accuracy.Xgboost random forest classifier (XGBRF) gives us random forest functionality [7]which is also called (bagging+ boosting) techniques,thatprovides the maximum accuracy of 99% and most of the real world problem can be solved using this technique.Using this simplest technique people can get the accurate result.

Xgboost random forest classifier (XGBRF) is actually the enhanced version of Xgboost classifier that trains random forest instead of gradient boosting used directly by Xgboost classifier and contain default values and some parameters which can be adjusted according to the need of algorithm in terms of accuracy.

The main parameter of XGBRF classifiers are learning rate, number estimators, booster, random state, sample by node, cole sample by node, subsample, base score which plays an important role in increasing its accuracy by playing with these parameters.

The value estimation is done below -

N_estimator-This parameter gives a number of tree to be trained

Learning rate- By default it is taken as 1

Subsample and cole_sampleby node- By default it is taken as 0.8

Booster-This parameter will take always gbtree

5. Performance evaluation

For the purpose of effective evaluation and experimentation results, the following parameters are used which are discuss below:

5.1 Confusion matrix

The concept of confusion matrix ideally Changes with respect to binary class classification and with respect to multiclass classification. Confusion metrics is basically a matrix which is prepared once our model is created. Using the confusion matrix we will calculate all four metrics.

5.2 Accuracy

The performance of any model can be very easily measured with the help of the accuracy parameter. Accuracy is simply the ratio of the total number of correct predictions to the total number of samples. The formula for accuracy is given below:

$$Accuracy - \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy lies between 0 to 1 accuracy of 0 is considered to be a bad model and accuracy of 1 is considered to be better model in my case accuracy is coming 99.12%

5.3 Precision and Recall

Precision and recall are generally use in information extraction problem. Precision is calculated for each label technically we have precision of malignant and precision of benign.it is defined as of all the data points the model predicted is positive and what % of them are actually positive in precision we don't have to care about negative class. the precision score for benign and malignant case is 98% in both the cases.

$$Precision = \frac{TP}{TP + FP}$$

Recall is all about true positive rate and recall is also caring about positive class rather than negative class .the recall score for benign and malignant case is 98% in both the cases.

$$Recall - \frac{TP}{TP + FN}$$

5.4 F1 SCORE

F1 score is the harmonic mean of precision and recall

F1 score = 2* Precision*recall/ Precision + recall

Where TP= True positive, TN= True Negative, FP= False positive and FN= False Negative

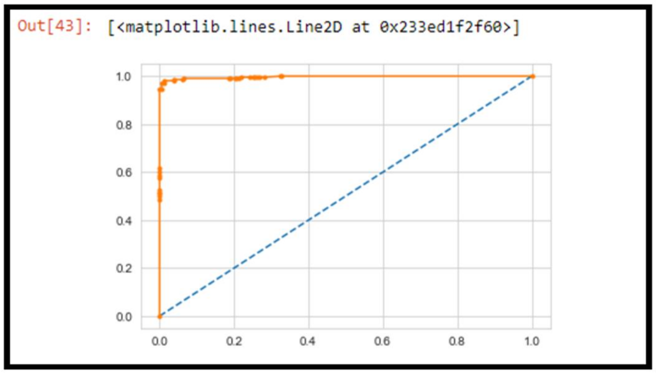
5.5 ROC - AUC CURVE

There is another matrix used in binary classification is called receiver operating characteristic curve and area under this curve. This curve was used by electronics and radio engineers during second world year to predict the working of missiles. ROC-AUC curve is a simple two dimensional curve in which x axis is false positive rate(FPR) and y axis is a true positive rate(TPR) where as FPR and TPR are given below:

FPR = FALSE POSITIVE/TRUE NEGATIVE+FALSE POSITIVE

TPR = TRUE POSITIVE/ FALSE NEGATIVE+TRUE POSITIVE

Model is good if the value of AUC is between 0.5 to 1.



```
Out[43]: [<matplotlib.lines.Line2D at 0x233ed1f2f60>]
```

Figure 1: ROC AUC Curve

6. CLASSIFICATION REPORT

The report shown in table 1 shows the real time evaluation of dataset [4], indicating the classification and the precision of algorithm. The area of cancer detection is very agile and comprehensive but the Xgboost approach to the solution creates a new impact. If we consider the classification average of Benign then it can be seen an F-1 score of 99 is observed making it the highest precision in the field. The works by the authors of [11] have paved some new features which are mentioned at Advanced levels.

Table 1 – Classification Report

Classification rates/type of breast cancer	Precision	Recall	F1-score	Support
Benign	98	99	99	357
Malignant	98	97	98	212
Micro average	98	98	98	569
Macro average	98	98	98	569
Weighted average	98	98	98	569

7. CONCLUSION

This paper presents the use of machine learning algorithms using ensemble learning for the diagnosis of breast cancer in shortest time when deployed on various machines at the time of deployment it requires only the different features which are responsible for the type of cancer,the precision and recall are used there which are two major metrics for the performance of the model which shows that model is sensible.Later Brute force technique is applied to get the highest accuracy and generalized model and we got the generalized model with maximum accuracy by changing its random state using short iterations & finally the deployment phase gave the correct prediction when checking on any new dataset with Image.

REFERENCES

1. World Health Organization, “Cancer country profiles 2014,” WHO, http://www.who.int/cancer/country-profiles/en/
2. .American cancer society https://www.fightcancer.org/breast-cancer
3. .National cancer institutehttps://www.cancer.gov/types/breast
4. Ucirepository (university of wiscosin)https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
5. Sklearnscikit learn https://scikit-learn.org/stable/
6. seaborn and counplothttps://jakevdp.github.io/PythonDataScienceHandbook/04.14-visualization-with-seaborn.html
7. XGBOOST RANDOM FOREST CLASSIFIER https://xgboost.readthedocs.io/en/latest/tutorials/rf.html
8. Roc auc curve https://en.wikipedia.org/wiki/Receiver_operating_characteristichttps://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.
9. L.G. Ahmad, A.T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi and A.R. Razavi, “Using three
10. machine learning techniques for predicting breast cancer recurrence,” (2013), J Health Med Inform
11. 4: 124. doi:10.4172/2157-7420.1000124.
12. Zahra Nematzadeh, Roliana Ibrahim and Ali Selamat, “Comparative studies on breast cancer
13. classifications with k-fold cross validations using machine learning techniques,” Proc. in 2015
14. 10th Asian Control Conf. (ASCC), pp 1-6, IEEE, 2015.
15. H. Hasan and N. M. Tahir, "Feature selection of breast cancer based on principal component
16. analysis," in Signal Processing and Its Applications (CSPA), 2010 6th International Colloquium
17. on, 2010, pp. 1-4.
18. Uma Ojha and Savita Goel, “A study on prediction of breast cancer recurrence using data mining
19. techniques,” 2017 7th Int. Conf. on Cloud Computing, Data Science & Engineering – Confluence,
20. pp 527-530, IEEE, 2017.
21. Soumadip Ghosh, Sujoy Mondal and Bhaskar Ghosh, “A comparative study of breast cancer
22. detection based on SVM and MLP BPN classifier,” 2014 1st Int. Conf. on Automation, Control,
23. Energy and Systems (ACES), pp 1-4, IEEE, 2014.
24. AlirezaOsareh and BitaShadgar. “Machine learning techniques to diagnose breast cancer,” 2010
25. 5th Int. Symp. on Health Informatics and Bioinformatics, 20-23 April 2010, Antalya, Turkey.
26. Dana Bazazeh and RaedShubair. “Comparative study of machine learning algorithms for breast
27. cancer detection and diagnosis,” 2016 5th Int. Conf. on Electronic Devices, Systems and
28. Applications (ICEDSA), 6-8 December 2016, Ras Al Khaimah, UAE.
29. Muhammad Sufyian Bin MohdAzmi and Z. C. Cob, “Breast cancer prediction based on
30. backpropagation algorithm,” Proc. of 2010 IEEE Student Conf. on Research and Development
31. (SCOReD 2010), pp 164-168, 13 - 14 Dec 2010, Putrajaya, Malaysia.
32. B. M. Gayathri and C. P. Sumathi, “Comparative study of relevance vector machine with various
33. machine learning techniques used for detecting breast cancer,” 2016 IEEE Int. Conf. on
34. Computational Intelligence and Computing Research (ICCIC), pp 1-5, IEEE, 2016.

35. Deepshri Hebbalkar, Smita Naik, Anuj Divkar, Avinkumar Pednekar Valerie Menezes & Shreedatta Sawant, “Breast Cancer Image Segmentation using Ekstrap and FCS Algorithm”, International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), Vol. 7, Issue 3, pp. 39-48
36. Basavaraj Hiremath & SC Prasannakumar, “Automated Evaluation of Breast Cancer Detection Using SVM Classifier”, International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), Vol. 5, Issue 1, pp. 7-16
37. Swetha Danthala, Seeramsrinivasa Rao, Kasiprasad Mannepalli & Dhantala Shilpa, “Robotic Manipulator Control by Using Machine Learning Algorithms: A Review”, International Journal of Mechanical and Production Engineering Research and Development (IJMPERD), Vol. 8, Issue 5, pp. 305-310
38. Duong Hung Bui, Manh Cuong Nguyen, Thi Hong Nguyen & Xuan Tho Dang, “Improve Efficiency of Cancer Classification by Combining Selected Feature and Additional Elements”, International Journal of Electrical and Electronics Engineering (IJEEE), Vol. 7, Issue 4, pp. 1-8
39. Prajna Das, Amit Kumar Adhya & Urmila Senapati, “Correlation of Estrogen Receptor(ER), Progesterone Receptor (PR) and Human Epidermal Growth Factor Receptor (HER-2/neu) Status with Histological Types, Grades and Stages of Breast Carcinoma”, International Journal of General Medicine and Pharmacy (IJGMP), Vol. 3, Issue 4, pp. 41-50
40. H. C. Nagaraj, Prasanna Paga & Kamal Lamichhane, “Early Breast Cancer Detection Using Statistical Parameters”, IMPACT: International Journal of Research in Engineering & Technology (IMPACT: IJRET), Vol. 2, Issue 3, pp. 31-