International Journal of Advances in Engineering and Management (IJAEM)
Volume 2, Issue 5, pp: 07-10 www.ijaem.net ISSN: 2395-5252

A Review on an Efficient Prediction of Breast Cancer Using Data Mining Techniques

Jahnavi Vyas[1], Ankit Kalariya[2], Nirali Borad[3]
Atmiya University, Rajkot

Date of Submission: 05-08-2020 Date of Acceptance: 22-08-2020

ABSTRACT: Breast cancer is the dangerous problem faced by many women. To predict and identify the tumor at earlier stage and to cure the cancer earlier is the most important need at this time. The tumor may be benign or malignant. Benign tumor means not so harmful but malignant tumor is considered as dangerous. So, the early diagnosis of the disease helps to prevent the cancer. There are different data mining algorithms used to predict the breast cancer at earlier stage. The main aim of this review paper is to review on various data mining techniques that are considered for an efficient breast cancer prediction.

Keywords: Breast cancer, Prediction, Data mining, Random Forests (RF), Naïve Bayes (NB), Decision Trees (DT), Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbor (KNN), Ensemble Voting Classifier, Multilayer Perceptron (MLP), Artificial Neural Network (ANN), Back Propagation.

I. INTRODUCTION

Breast cancer is one of the most dangerous diseases in women. The human body is composites of millions of cells. When the irregular growth of cell starts then it form a tumor which leads to the cancer.

The survey of 2013 indicates that 230,815 ladies and 2109 gents in the US were determined to have breast cancer [3]. In the study, confirmed the forecasted increase in U.S. breast cancer cases diagnosed each year, concluding that they will grow from 283000 cases in 2011 to 441000 in 2030 [10]. Breast cancer diagnosis and prognosis are two medical applications of data mining that pose a great challenge to the research community [1]. Deep learning analysis of Wisconsin Breast Cancer Database used for detection and classification of breast cancer by applying Naïve Bayes, Decision Trees, Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbor (KNN), Random Forest (RF) and Multi Layer Perceptron (MLP) algorithms, higher accuracy is obtained which is up to 98% to 99% [2].

Machine learning helps us to extract information and knowledge from this the basis of past experiences and detect hard-to-perceive pattern from large and noisy dataset [4]. Wisconsin Prognosis Breast Cancer (WPBC) dataset is widely used and that is publically available from the university of Wisconsin Hospitals. Having a limited set of classes, certainly including the breast cancer sample, it is easier to identify the right class, and the result would be more accurate than with clustering algorithms [5]. For determine correctly weather a cancer is benign or malignant require consideration of doctor’s experience and patient’s clinical reports and many other factors [6]. To minimize the over fitting problem, we used 10-fold cross validation technique which splits the data into 90% for training and 10% for validation test for each baseline model and the process is repeated a few times until we get a stable accuracy score from the average 10-fold [7]. After the features were extracted and selected, we can apply the machine learning methods to the data that we obtained. The machine learning methods to be applied, as discussed previously, are K-Nearest Neighbors, Support Vector Machines, Naive Bayes and Random Forest [8].

The ground methodology on which random forests is based is recursion [9]. RF is very efficient with large datasets and high dimensional Data [11]. The k-Nearest-Neighbors (KNN) is a technique which is used for the classification of data in machine learning; it will perform classification by finding the nearest and similar

Data points within the corresponding dataset [12].

Data mining approach has helped a lot in medical science due to its high efficiency in the prediction of the future health condition, and also helps in reduction of medical cost and improving the health of people and quality in real time which helps in saving lives’ of people [13].With the advancement of technology and machine learning techniques, the cancer diagnosis and detection accuracy has improved and provide the better result. Now we look at some data mining methods by which prediction of breast cancer became easy.

II. REVIEW PAPER

[1] D. A. Almuhaidib et al. (2018): In this paper, Wisconsin Prognostic Breast Cancer (WPBC) dataset is used that is publically available in the UC Irvine Machine Learning Repository. The dataset has 33 features and total 198 instances with 148 non-recur cases and 46 recur cases. The dataset has 4 missing values. First we preprocess the dataset, the preprocessing steps performed before building the model are feature selection, ranking and extraction. For feature selection recursive feature elimination is used, for feature ranking random forest is used and for extraction principle component analysis is used. After preprocessing is done, 10-fold cross-validation technique is used to split the dataset. The system works in two layers. In the first layer, training and testing of dataset is conducted using different algorithms and continue training and testing process until come up with the right model. The right model is the best performing model. Here three models are observed as best models-Random Forest, Decision Tree and Naïve Bayes. Random Forest performs best. Average of these three models is also performed. At the end right model is deployed. In second layer, best performing model is used in website for real time prediction. Highest accuracy is achieved by Random Forest (65.22%)

[2] S. Ghosh et al. (2017): The breast cancer dataset of the university of Wisconsin Hospitals is used in this experiment. The dataset has 699 instances but some of instances are deleted due to missing attributes. Each instance has two possibilities that are Benign or Malignant. Further data is analyzed and come up with total 30 attributes and 569 instances. The algorithms used in this project are Naïve Bayes, Random Forest, K-Nearest Neighborhood, Support Vector Machine, Logistic Regression, and Decision Tree. Also Neural Network and CNN classifiers are used. The mean values of cell radius, perimeter, area, compactness, concavity and concave points can be used in classification of the cancer. Larger values of these parameters show a correlation with malignant tumors. The dataset is split into train data and test data. Above described algorithms are used for binary classification and their accuracy of detection are compared so that the appropriate model can be found. Weka machine learning tool is used in the experiment. The highest accuracy is obtained from Random Forest. With the deep networks, convergence time increase and it get harder to optimize the network. Random Forest performs best with the accuracy score 96.83%.

[3] R. D. Ghongade et al. (2017): In this research, Computer-Aided Diagnosis system is used for the detection of abnormality in mammograms. The digital mammogram images with 1024*1024 pixels are imported from MIAS database. Main issues for extracting features of mammographic images are noise, different resolution, low quality and contrast. This make makes the location of tumor harder. Pre-processing is done to conquer these issues and to make efficient feature extraction of image. Firstly, the Gaussian filter is applied for smoothing the image. Gaussian blur is used to blur the image and remove noise from image. Then histogram equalization is used to enhance the contrast of grayscale image. After pre-processing, Region based segmentation is used to segment masses from its background and then it is multiplied with original image as the normalization. After segmentation, some features are extracted using gray level co-occurrence matrix. Then some relevant features are selected using FCBF technique. After that the classification is done using Random Forest classifier. The high accuracy is achieved by Random Forest is 97.32%.

[4] M. Gupta et al. (2018): In this paper, a comparative analysis of four machine learning techniques is performed on Wisconsin breast cancer database to predict the breast cancer recurrence. The techniques used are Multilayer perceptron, Decision Tree, Support Vector machine and K-Nearest Neighbor. The dataset has 569 samples and 32 attributes. The dataset contains non-missing attribute and also contains 212 malignant and 357 benign attributes. Firstly, the collected dataset is separated for training and testing data (70-30%). Machine is trained using training data on the basis of classis correspond to samples. In pre-processing stage, label and id columns are removed and mean, standard deviation of each attribute is calculated. Then in feature extraction, efficient features are processed to train the machine. Examine mean of each attribute and settles each attribute in increasing order on the basis of difference between means of attribute and then accuracy performs on all attributes. A machine learning model is generated by above mentioned classifiers for dataset. Data is tested which is used to simulate the constructed model by classifying data into cancerous tumor and non-cancerous tumor. Multilayer perceptron performs best with the accuracy score 98.12%.

[5] Y. Khourdifi et al. (2018): This research uses publically available dataset from the University of Wisconsin Hospitals. The dataset has total 699 numbers of cases but some instances are deleted due to missing attributes. There are two possibilities for each instance: Benign or Malignant. After further analysis, the dataset has 569 instances and 30 attributes. After the features are extracted and selected, the machine learning methods like K-Nearest Neighbor, Support Vector Machine, Naïve Bayes and Random Forest are applied. 10-fold cross validation is also used to evaluate the models that divide the set in training sample to form the model and a set of tests to evaluate it. After applying the pre-processing and methods, the dataset is visually analyzed and determined the distribution of values in terms of effectiveness and efficiency. The effectiveness of all classifiers is evaluated in terms of time to build the model, correctly and incorrectly classified instances and accuracy. At the end of experiment, the conclusion is that SVM is the best classifier than others. SVM has the accuracy 97.9%.

[6] B. Liu et al. (2018): In order to improve the accuracy, speed of prediction and to obtain the most relevant features four feature selection methods are used: feature selection based on relevance, the recursive feature elimination, recursive feature elimination cross validation and principle component analysis. The most important features are texture_mean, area_mean, concavity_mean, area_se and concavity_worst. Data are collected from digitized images of a fine needle aspirate (FNA) of breast mass. They describe characteristics of the cell nuclei presented in the image. The dataset has total 569 instances including 357 benign and 212 malignant cases. Used classification methods are: Support Vector Machine, Decision Tree and Random Forest. Among these machine learning methods Random Forest performed best with the accuracy score 99%.

[7] Q. H. Nguyen et al. (2019): In this paper, Wisconsin Breast Cancer Dataset (WBCD) is taken from the UCI Machine learning repository. The dataset contained 569 instances taken from needle aspirates from patients' breast. In the dataset, 357 cases were Benign and 212 cases were Malignant. The dataset was analyzed and pre-processed by the steps: missing value checking, class imbalance checking, normalization checking, correlation checking and train/test split.

The data is split into training and testing sets in the ratio of 70:30 respectively. The benchmark model is created using Random Forest classifier. Various models are trained and tested on the taken data after feature scaling and principle component analysis. Cross validation is performed which showed that model is stable. Among all the evaluated models only few models: Ensemble-Voting classifier, Logistic Regression and Support Vector Machine returned with the accuracy of at least 98%.

III. LITERATURE SURVEY

Author	Year	Dataset	Methods
D. A. Almuhaidib et al.	2018	Wisconsin	RF, NB, DT
S. Ghosh et al.	2017	Wisconsin	NB, RF, KNN, SVM, LR
R.D. Ghongade et al.	2017	MIAS dataset	RF, SVM
M. Gupta et al.	2018	Wisconsin	RF, KNN, SVM, DT, MLP
Y. Khourdifi et al.	2018	Wisconsin	RF, KNN, SVM, NB
B. Liu et al.	2018	Digitized image of Fine Needle Aspirates (FNA) of breast mass	RF, DT, SVM
Q. H. Nguyen et al.	2019	Wisconsin	Ensemble Voting, LR, SVM, RF
Y. Khourdifi et al.	2018	Wisconsin	KNN, SVM, RF, NB, MLP
S. Sharma et al.	2018	Wisconsin	RF, KNN, NB
R. D. Ghongade et al.	2017	MIAS dataset	RF, RF-ELM
M. S. Yarabarla et al.	2019	Wisconsin	KNN, SVM, RF

P. Singhal et al.	2018	Wisconsin	ANN, Back Propagation

IV. CONCLUSION

In this review paper, we discussed about various data mining techniques by which we can predict breast cancer. Different models might carry different knowledge about the data, which hopefully produce more stable model when combined. The accuracy depends upon pre-processing methods and different data mining algorithms by which the models are created. From the above study, we can infer that there is a still lack of early diagnosis, accuracy, sensitivity and specificity of the breast cancer data. So by analyzing data and algorithms we can get more accurate result.

REFFERENCE

[1]. D. A. Almuhaidib et al., "Ensemble Learning Method for the Prediction of Breast Cancer Recurrence," 2018 1st International Conference on Computer Applications & Information Security (ICCAIS), Riyadh, 2018, pp. 1-6.

[2]. C. Shahnaz, J. Hossain, S. A. Fattah, S. Ghosh and A. I. Khan, "Efficient approaches for accuracy improvement of breast cancer classification using wisconsin database," 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Dhaka, 2017, pp. 792-797.

[3]. R. D. Ghongade and D. G. Wakde, "Computer-aided diagnosis system for breast cancer using RF classifier," 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, 2017, pp. 1068-1072.

[4]. M. Gupta and B. Gupta, "A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques," 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), Erode, 2018, pp. 997-1002.

[5]. Y. Khourdifi and M. Bahaj, "Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification," 2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), Kenitra, 2018, pp. 1-5.

[6]. B. Liu et al., "Comparison of Machine Learning Classifiers for Breast Cancer Diagnosis Based on Feature Selection," 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan, 2018, pp. 4399-4404.

[7]. Q. H. Nguyen et al., "Breast Cancer Prediction using Feature Selection and Ensemble Voting," 2019 International Conference on System Science and Engineering (ICSSE), Dong Hoi, Vietnam, 2018, pp. 250-254.

[8]. Y. Khourdifi and M. Bahaj, "Feature Selection with Fast Correlation-Based Filter for Breast Cancer Prediction and Classification Using Machine Learning Algorithms," 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Rabat, Morocco, 2018, pp. 1-6, doi: 10.1109/ISAECT.2018.8618688.

[9]. S. Sharma, A. Aggarwal and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), Belgaum, India, 2018, pp. 114-118, doi: 10.1109/CTEMS.2018.8769187.

[10]. "Study forecasts new breast cancer cases by 2030," April 23, 2015, By NCI Staff, National Cancer Institute at the National Institutes of Health, USA.

[11]. R. D. Ghongade and D. G. Wakde, "Detection and classification of breast cancer from digital mammograms using RF and RF-ELM algorithm," 2017 1st International Conference on Electronics, Materials Engineering and Nano-Technology (IEMENTech), Kolkata, 2017, pp.1-6, doi:10.1109/IEMENTECH.2017.8076982.

[12]. M. S. Yarabarla, L. K. Ravi and A. Sivasangari, "Breast Cancer Prediction via Machine Learning," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019,pp.121124,doi:10.1109/ICOEI.2019.8862533.

[13]. P. Singhal and S. Pareek, "Artificial Neural Network for Prediction of Breast Cancer," 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on, Palladam, India, 2018, pp. 464468,doi:10.1109/ISMAC.2018.8653700.