East West University

Project Report

Breast Cancer Prediction

Submited by

Rahul Shaha	2017-2-60-103
Md Ali	2017-2-60-045
Sidratul Moontha	2017-2-60-071
Mijanur Rahman	2017-2-60-070

Submitted to,

Md Mahmudur Rahman

Department of Computer Science & Engineering

Submission Date

26/May/2021

Introduction:

World Health Organization (WHO) reported the breast cancer is the most common cancer amongst death world. Breast cancer is the most found disease in the women, worldwide, where abnormal growth of a mass of tissue, cause the expansion of malignant cells leads to acute breast cancer. These malignant cells are originally created from milk glands of the breast. These malignant cells which are the main reason for breast cancer can be classified into different groups according to their unusual progress and capability affecting other normal cells. The capability of affecting means whether these malignant cells affect only the local cells or can spread throughout the full body. The effect of spreading these malignant cells throughout the whole body of the patient is called as metastasis [7]. It is very important to prevent this spreading effect by a diagnosis of cancer in the early stages using advanced techniques and equipment.

Objectives:

The objective of this paper will be predictive analysis of breast cancer using various machine learning techniques like Naïve Bayes method, Linear Discriminant Analysis, K-Nearest Neighbors and Support Vector Machine method.This research aims to find alternative methods which are easier to implement and work with different data and to prevent the malignant cells spreading effect by diagnosis of cancer by early stage using machine learning algorithm.

Motivations:

Many people are affected from breast cancer at the present time. Causing of this diseasedepends on man factors and cannot be simply determined. In addition, the identification method that determines whether or not the cancer is benign or malignant additionally needs an excellent deal of effort from a doctors and physicians. Once many tests are concerned within the identification of breast cancer, like clump thickness, uniformity of cell size,uniformity of cell form, etc., the ultimate result could also be troublesome to get, even for doctors. This has given an increase within

the previous few years to the utilization of machine learning and computing generally as diagnostic tools. The diseases that take numerous lives,diagnostic computer-based applications are used wide.

We aim in this project to compare different classification learning algorithms sig-nificantly to predict a benign from malignant cancer in breast cancer dataset. We do this project to investigate different machine learning techniques and we will use several algorithms andapply on breast cancer dataset. We will focus on machine learning algorithms: Naïve bayes,K-nearest neighbor, logistic regression, reinforcement algorithm, support vector machinealgorithm. We will primarily study these various algorithms and analyze their result.

Existing work:

Several studies have been conducted on the implementation of ML on Breast Cancer detection and diagnosis using different methods or combination of several algorithms to increase the accuracy. S. Gc *et al.* [17] worked on extracting features including variance, range, and compactness. They used SVM classification to evaluate the performance. Their findings showed the highest variance of 95%, range 94%, compactness 86%. According to their results, SVM can be considered as an appropriate method for Breast Cancer Detection.

Wang and Yoon [22] chose four methods of Data Mining to measure their effectiveness in detection. These models were: SVM, ANN, Naïve Bayes Classification and Adaboost tree. In addition, PCs and PCi were used for making hybrid models. After checking the accuracy, they have found out that Principal Component Analysis (PCA) can be a critical factor to improve performance.
Hafizah *et al.* [23] compared SVM and ANN using four different datasets of breast and liver cancer including WBCD, BUPA JNC, Data, Ovarian. The researchers have demonstrated that both methods are having high performance but still, SVM was better than ANN.
Azar and El-Said [24] worked on six different methods of SVM. They have compared ST-SVM with LPSVM, LSVM, SSVM, PSVM, and NSVM to find out which method performs the best in accuracy, sensitivity, specificity, and ROC. LPSVM proved to be the best with accuracy 97.1429%, sensitivity 98.2456%, specificity 95.082%, and ROC 99.38%. Therefore, LPSVM has the highest performance.

Necessity: he breasts cancer prediction using machine learning Here we can easily identify the patient's disease. In this way patients can be easily diagnosed.

Data Set:

Data set collect form kaggle . There are 578 instance and 30 input attribute and 1traget attribute. fig1 shown that part of the data set.

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	radius error	texture error	perimeter error	area error
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	0.2419	0.07871	1.0950	0.9053	8.589	153.40
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	0.2069	0.05999	0.7456	0.7869	4.585	94.03
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	0.2597	0.09744	0.4956	1.1560	3.445	27.23
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	0.1809	0.05883	0.7572	0.7813	5.438	94.44
564	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	0.1726	0.05623	1.1760	1.2560	7.673	158.70
565	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	0.1752	0.05533	0.7655	2.4630	5.203	99.04
566	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	0.1590	0.05648	0.4564	1.0750	3.425	48.55
567	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	0.2397	0.07016	0.7260	1.5950	5.772	86.22
568	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	0.1587	0.05884	0.3857	1.4280	2.548	19.15

Fig1: Data set

Data describe:

Fig2 Data set show to data count, mean, std, min, percentage and maximum value.

	count	mean	std	min	25%	50%	75%	max
mean radius	569.0	14.127292	3.524049	6.981000	11.700000	13.370000	15.780000	28.11000
mean texture	569.0	19.289649	4.301036	9.710000	16.170000	18.840000	21.800000	39.28000
mean perimeter	569.0	91.969033	24.298981	43.790000	75.170000	86.240000	104.100000	188.50000
mean area	569.0	654.889104	351.914129	143.500000	420.300000	551.100000	782.700000	2501.00000
mean smoothness	569.0	0.096360	0.014064	0.052630	0.086370	0.095870	0.105300	0.16340
mean compactness	569.0	0.104341	0.052813	0.019380	0.064920	0.092630	0.130400	0.34540
mean concavity	569.0	0.088799	0.079720	0.000000	0.029560	0.061540	0.130700	0.42680
nean concave points	569.0	0.048919	0.038803	0.000000	0.020310	0.033500	0.074000	0.20120
mean symmetry	569.0	0.181162	0.027414	0.106000	0.161900	0.179200	0.195700	0.30400
ean fractal dimension	569.0	0.062798	0.007060	0.049960	0.057700	0.061540	0.066120	0.09744
radius error	569.0	0.405172	0.277313	0.111500	0.232400	0.324200	0.478900	2.87300

Fig:2

Methodology:

With this project model I can see if any patient has been infected with this disease. With this machine learning model, the patient can be compared with the doctor's prediction. Different types of Classifier Algorithm have been used in this model.

Support Vector Classifier:

Support vector machine (SVM) provides accurate classification but suffers from a large amount of computation. The accuracy score of SVC is 98%.SVC gives us 98% correct detection of those people who have actually breast cancers provides incorrect detection which is only 2%. So we can say that SVC classifier is one of the suitable algorithm in detecting breast cancer.

The results obtained from the wide range of benchmark problems show that the OSVC algorithm has a much faster convergence and results in a smaller number of support vectors for the same quality of pattern classification and a better generalization performance in comparison with the existing algorithms.

Logistic Regression:

Logistic regression examines the relationship between a binary outcome (dependent) variable such as presence or absence of disease and predictor (explanatory or independent) variables such as patient demographics or imaging findings. In this project we found that Logistic Regression algorithm gives us 95% accuracy and 5% incorrect result in breast cancer detecting method.

So, SVC is much better than Logistic Regression algorithm.

KNN Classifier:

K-NN is relatively simple and effective classification algorithm when compared to other algorithms , it is non-pragmatic algorithm does not need the assumption for distributing data . It classifies the case study directly by the samples in the data set and thus, does not require a training process. It's an easy supervised learning algorithm for pattern recognition.

KNN algorithm shows 96.49% accuracy and 3.51% inaccuracy.

Naive bayes classifier:

The Naive Bayes algorithm is a simple probability classifier that calculates a set of probability by counting the frequency and combinations of values in a given data set. The algorithm uses Bayes ‘theorem and assumes that all variables are independent considering the value of the class variables.

The accuracy of Naive bayes classifier is 93% and inaccuracy is 7%.

Decision Tree Classifier:

In the data mining community, decision tree algorithms are very popular since they are relatively fast to train and In data mining, decision tree algorithms are very popular due to their characteristics such as fast to train and produce transparent models.

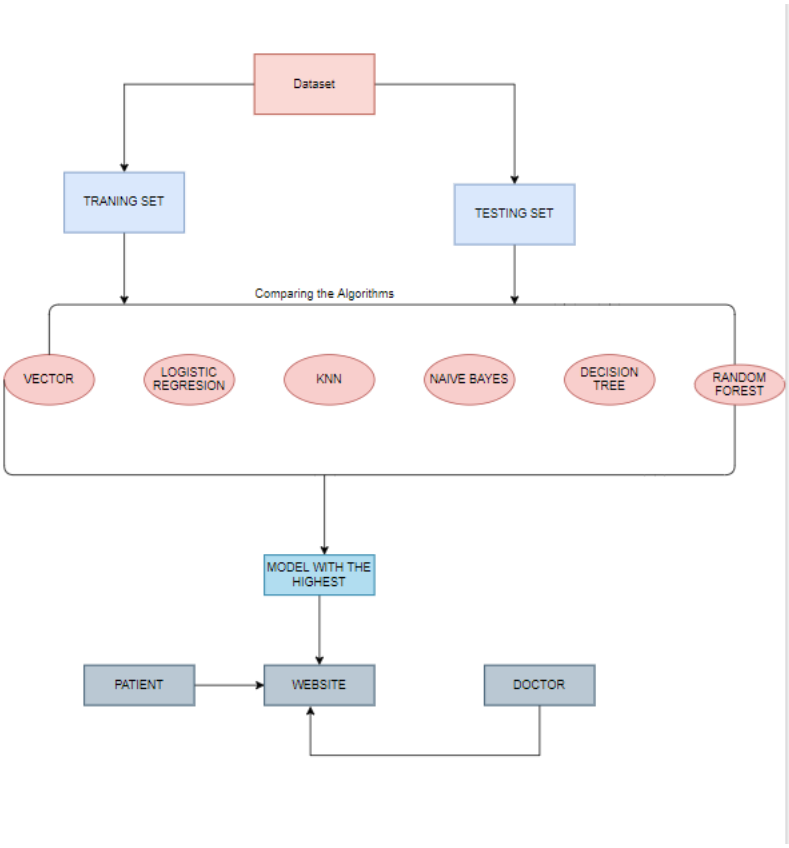
The accuracy level is 93.8% and inaccuracy level is 6.2%.

Random Forest Classifier:

On these datasets we obtained classification accuracy of 98.24% in the best case and of around 99.8% on average. This is very promising compared to the previously reported results.

It is observed that RF classifier and SVC are best among them. We can use SVC or RF for achieving the best result in breast cancer detection.

Block diagram:



Summary of prediction:

algorithm	AC
Vector classifier	98.25%
Logistic regression	95.61%
KNN classifier	96.41%
Naïve	93.86%

Bayes Classifier	
Decision tree classifier	93.91%
Random forest classifier	98.24%

Result: In this research, we are following classification algorithms workflow to process diabetes data set to create prediction model. Performance of the algorithms is summarized in Table to accuracy by algorithm.

Conclusion:

Challenges: In this report we have reviewed different machine learning algorithms for the prediction of breast cancer. Our main focus is to find out the most suitable algorithm that can predict the occurrences of breast cancer more effectively. The main purpose of this review is to highlight all the previous studies of machine learning algorithms that are being used for breast cancer prediction, this report provides the all-necessary information to the beginners who want to analyze the machine learning algorithms to gain the base of deep learning. The review of this

report is started from the types of breast cancer, fourteen research papers have been reviewed to get some knowledge about the major types, symptoms and causes of breast cancer. After that, the review of major machine learning techniques has been provided and this technique deeply elaborate algorithms that are being used for the predictions of breast cancer.

Limitations: While we were successful at attaining results with precise accuracies, there were certain hindrance which build up while carrying out this thesis. The initial issue was the lack of a significantly large dataset. While we did achieve accuracy with over 90% without PCA for all algorithms except decision tree, it cannot be denied that the algorithms could have been tested better with a large dataset. Availability of a large dataset could also test the runtime of algorithms run with PCA, since it is very difficult to trace out exactly how fast the algorithms run after PCA is applied on current dataset. Furthermore, there is a lack of complex models used in this thesis. Even though we obtained better results with the models we used, use of more complex models can capture complex interactions among features.

Future directions: Despite attaining accurate results and accuracies with the algorithms we have used; we wish to confirm the results we obtained are not biased thanks to the scale of our dataset. We would like to search out an even bigger dataset and perform similar analysis and see if the results are the identical. Furthermore, since our dataset is kind of obsolete, more criteria for prediction and improved technology must have been available to attain more accurate numerical data. It would also put our analysis to the test, if we can identify the right parameters from our current and future datasets in order to generate ROC curves. Additionally, besides the models we have tried, we would conjointly wish to attempt other algorithms such as Ada boost in order to compare results and continue our search for the best model for prediction. The idea of applying other feature selection on the currently used models is also under consideration, such as the Recursive Feature Elimination and the Correlation Heat Map.

Reference:

https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

https://www.ijert.org/research/breast-cancer-classification-and-prediction-using-machine-learning-IJERTV9IS020280.pdf

https://app.diagrams.net/?rev=0B4b_77y1kLxAU1l3T0JpejdHZjVPUXVwMTRWTGhCbGd1N00wPQ&chrome=0&nav=1&layers=1&edit=_blank&page=-1&mode=google&gfw=1#G13Is42k9cMIuYbxtFSTaEu1cj7PP-0OMH

https://cordis.europa.eu/docs/results/246/246479/final1-namdiatream-final-publishable-report-.pdf