ISSN 2278-3091
Volume 9 No.2, March -April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse12922020.pdf
https://doi.org/10.30534/ijatcse/2020/12922020

Breast Cancer Detection Using Machine Learning

Karthikeyan B[1], Sujith Gollamudi[2],Harsha Vardhan Singamsetty[3], Pavan Kumar Gade[4], Sai Yeshwanth Mekala[5]

School Of Computing, SASTRA Deemed To Be University, Thanjavur-613401, India, mbalakarthi@gmail.com[1]
School Of Computing, SASTRA Deemed To Be University, Thanjavur-613401, India, sujithreddy727@gmail.com[2]
School Of Computing, SASTRA Deemed To Be University, Thanjavur-613401, India, harshasingamshetty@gmail.com[3]
School Of Computing, SASTRA Deemed To Be University, Thanjavur-613401, India, pavankumargade9@gmail.com[4]
School Of Computing, SASTRA Deemed To Be University ,Thanjavur-613401, India, saiyeshwanthscts1234@gmail.com[5]

ABSTRACT

Among different cancers, the most invasive and menacing cancer in women is breast cancer. Women in 140 of 184 countries across the world, continually gets affected by this. It accords for the majority of the cancer victims. The Death rate is around 13 minutes per person. So, earlier detection of this is a blessing in disguise because it increases the chances of successful treatment by providing care at the earliest possible stage. The above idea is the motive for this project. The implementation of this project is done by using Machine Learning techniques on data from the UCI Machine Learning Repository Data Set.

Key words: Machine Learning, Breast Cancer Detection, Logistic Regression, Decision Tree, Random Forest

1.INTRODUCTION

Breast Cancer is one of the cancers which is claiming women's lives.Breast cancers are categorized according to the area in which they begin, that is, the ducts, the lobules, and the tissue in the middle. They can however generally be classified as benign and malignant [13]. A Benign tumor that doesn't spread to the other breast tissues and Malignant tumors spread to the other breast tissues. Previously many Technologies have been used to identify breast cancer, By Mammography the breast tissue is screened which decreases the death rate of breast cancer, but it has certain limitations. Ultrasound Imaging is also used to identify by sending ultrasonic waves into the body, but it also has limitations like it cannot detect tumor which size is less than 5 mm. Sonography is another technology to identify the tumors initially it is done with mammography and Sonography checks any abnormalities in it. By Infrared Thermography, Infrared sensors are used to image the heat differences in the breast tissue. Regions which are having high temperature are considered as tumors [1]. All these technologies didn't show accurate results and were a kind of risky.

Nowadays Technology is well developed and new technologies were getting more accurate results than older ones. Machine learning is part of Artificial Intelligence(AI) which allows systems to learn automatically based on machine learning algorithms and improve the performance of the system without any programming. Doctors want to distinguish between these tumors through a good diagnosis process. Tumors even by specialists are generally very difficult to identify. Automation of the diagnostic device is also essential to diagnose tumors. Several researchers pursued early detection of cancer by machine learning techniques. Such approaches perform well in cancer diagnostics, as the researchers have shown [14]. Machine Learning techniques are the most used techniques for finding breast cancer and these techniques are easy and safe for the patients. These techniques include logistic regression, Random Forest, Decision Tree, Etc ., Logistic Regression uses a cancer database and classifies the tumors as benign and malignant based on the size and nature of tumor which gives 99.3 percent accuracy in the result. The random forest technique combines all the basic multiple algorithms and gives the result. It uses classification and regression methods to solve problems which gives 96.5 percent accuracy as a result. Decision Tree classifier method gives 93.71 percent accuracy. So These machine learning techniques give more accurate results when compare to older technologies which gives 70 to 80 percent accuracy.

2.LITERATURE SURVEY

Naresh Khuriwal et. Al. discussed ensemble machine learning algorithm for prediction of breast cancer, standardization method is used for pre-processing and Univariate feature selection algorithm for feature scaling and

has collected 16 features from data set and have achieved 98.50% accuracy[2].

Ebru Aydındag Bayrak et. Al. discussed Artificial Neural Network(ANN) and Support Vector Machine(SVM) algorithms for prediction of breast cancer in the WEKA tool and concluded that the Support vector machine has shown best performance compared to Artificial Neural networks based on performance metrics as Support vector machine has achieved 96.9957%accuracy[3].

Mohamed NEMISSI et. Al. proposed a classification system for the prediction of breast cancer which is based on an enhanced single hidden neural network which is trained using the Extreme Learning Machine(ELM) algorithm. sigmoid activation function has been used for hidden neurons[4].

Meriem Amrane et. Al. proposed two classifiers namely Naive Bayes classifier and K Nearest Neighbor classifier in which the KNN classifier gives high accuracy of 97.51% than the NB classifier which gives an accuracy of 96.19%[5].

Shubham Sharma et. Al. compares Random Forest, Naive Bayes, K-Nearest Neighbor algorithms for this comparison Wisconsin breast cancer data was used. These algorithms are compared in terms of accuracy and precision[6].

Mohammad Rasoul Al-Hadidi et. Al. proposed two methods namely Back Propagation Neural Network model and Logistic Regression model which uses mammography images as input. The results showed that the LR model utilizes more features than the BPNN model[7].

Mamatha Sai Yarabala et. Al. used Computer-Aided Detection or Diagnosis systems and Machine Learning Techniques to identify whether the person is having breast cancer or not with the help of trained data[8].

Noushin Jafarpisheh et. Al. predicted the recurrence of breast cancer by using a multi-layer perceptron. In this work, two different outputs are used namely Deep neural network and rough neural network. In addition to this support vector machine is also used[9].

Dana Bazazeh et. Al. compares three different Machine learning algorithms. They are Random Forest, Bayesian Networks and Support Vector Machines. Wisconsin original breast cancer data set is used as a training set and compared the performance based on accuracy, area of ROC curve, precision and recall[10].

Youness Khourdifi et. Al. compares K-Nearest Neighbors, Support Vector Machines, Random Forest and Naive Bayes. Support Vector Machines gave an accuracy of 97.9%. So, based on accuracy Support Vector Machines is selected[11].

Sri Hari Nallamala et. Al. Discussed breast cancer diagnostics by using the successful BCDNFIS approach by raising the complexities of the Sugeno and Fuzzy Inference systems in Mamdani [15].

3.RESULTS AND DISCUSSION

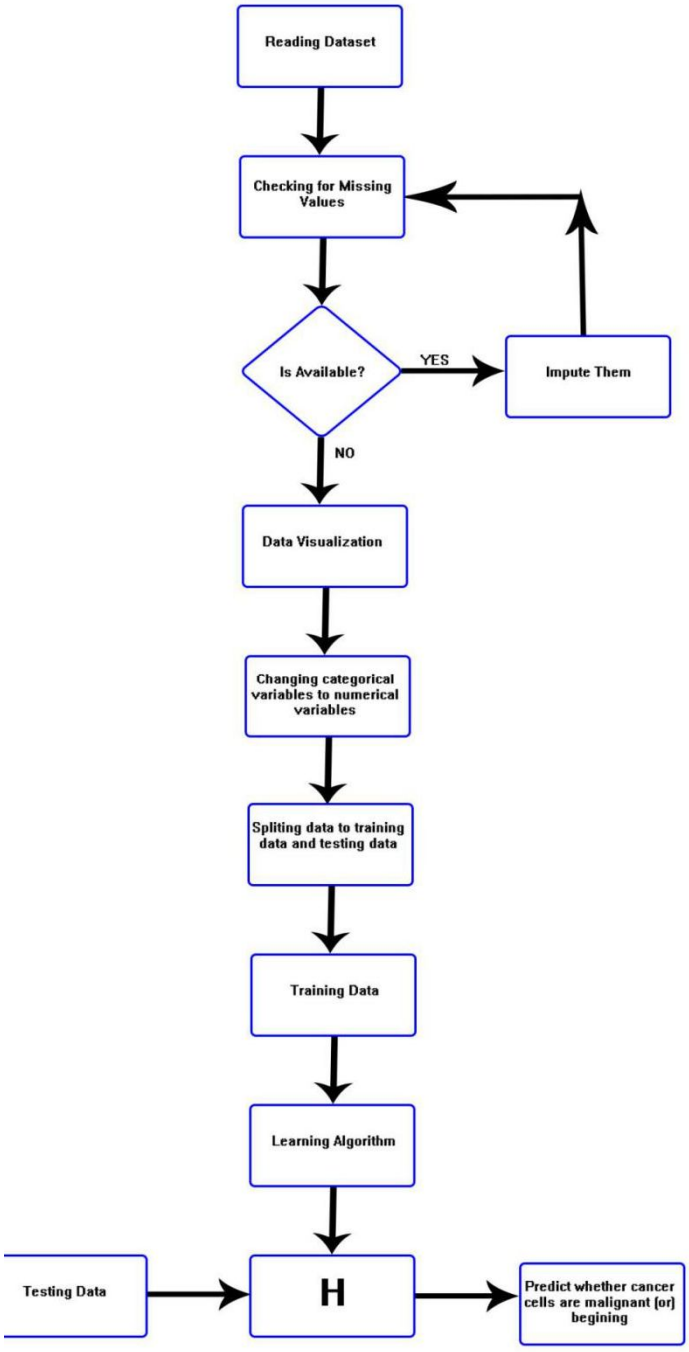

Figure 1: Flow Chart for Breast Cancer Detection

First, the Breast Cancer Wisconsin (Diagnostic) Data Set[12] collected from the UCI machine learning repository must be read. It consists of 569 rows and 32 columns. Then, check for missing values in the data set and impute if any. Next, visualize the data attributes using histograms, box plot, etc. Now, to work with continuous variables, the categorical variables in our data set are converted into continuous variables by performing one-hot encoding. Next, the data set is split into the training data set and testing data set. Then feature scaling is performed by using Standard Scaler. Standard scaler will normalize the features such that each feature will be having mean as 0 and standard deviation as 1. Then the training data set is given to learning algorithms like logistic regression, random forest, and decision tree to generate models. Now to predict whether the cancer cells are malignant or benign, the test data should be given to the model which was generated by using the training data set. Finally, for the Breast Cancer Wisconsin (Diagnostic) Data Set[12] Logistic regression has achieved 99.3% accuracy, Random Forest Classifier has achieved 96.5% accuracy and decision tree Classifier has achieved 93.7% accuracy. Flow chart for the Breast Cancer Detection has been shown in Figure 1. Confusion Matrix for Logistic Regression, Random Forest, Decision Tree Classifier are shown in Table 1, Table 2, Table 3 respectively. Evaluation Metrics for various Machine Learning Techniques used in this work are shown in Table 4 and Sample code for the Breast Cancer Detection is shown in Figure 2.

Table 1: Logistic regression confusion matrix

	Predicted Negative	Predicted Positive
Actual Negative	90	0
Actual Positive	1	52

Table 2: Random Forest Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	89	1
Actual Positive	4	49

Table 3: Decision Tree Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	88	2
Actual Positive	7	46

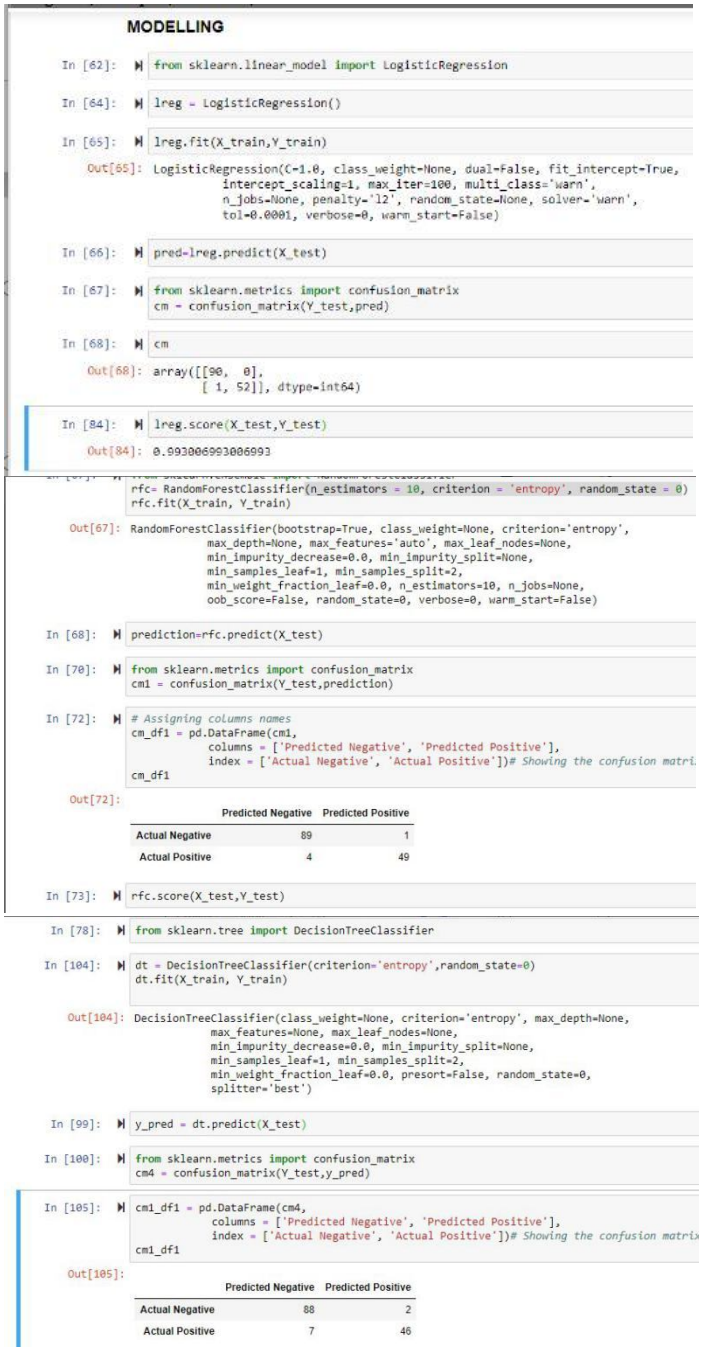

Figure 2: Sample code for Breast Cancer Detection

Table 4: Evaluation Metrics for various methods

Method used	Accuracy	Precision	Sensitivity	Specificity	F1 Score
Logistic Regression	0.9930	1.0	0.9811	1.0	0.9905
Random Forest	0.9650	0.98	0.9245	0.9889	0.9515
Decision Tree	0.9371	0.9583	0.8679	0.9778	0.9109

4.CONCLUSION

There is a chance of 12% for a woman picked randomly to be diagnosed with breast cancer. So the early detection of breast cancer will be very useful in treatment. In this work, various Machine learning supervised algorithms are used for the detection of Breast Cancer. On the Breast Cancer Wisconsin (Diagnostic) Data Set 3 main algorithms are used which are Logistic Regression, Random Forest Classifier, and Decision Tree Classifier. It has been observed that Logistic regression has achieved a higher efficiency of 99.3% accuracy whereas Random Forest Classifier has achieved 96.5% accuracy and Decision Tree Classifier has achieved 93.7% accuracy. Thus supervised machine learning algorithms will be very useful in early diagnosis and prognosis of cancer type.

REFERENCES

[1] Institute of Medicine (US) and National Research Council (US) Committee on New Approaches to Early Detection and Diagnosis of Breast Cancer; Joy JE, Penhoet EE, Petitti DB, editors. **Saving Women's Lives: Strategies for Improving Breast Cancer Detection and Diagnosis.** Washington (DC): National Academies Press (US); 2005. Appendix A, Breast Cancer Technology Overview. Available from: https://www.ncbi.nlm.nih.gov/books/NBK22310/

[2] N. Khuriwal and N. Mishra, "**Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm",** 2018 IEEMA Engineer Infinite Conference (eTechNxT), New Delhi, 2018, pp. 1-5.

[3] E. A. Bayrak, P. Kırcı and T. Ensari, "**Comparison of Machine Learning Methods for Breast Cancer Diagnosis",** 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), Istanbul, Turkey, 2019, pp. 1-3.

[4] M. NEMISSI, H. SALAH and H. SERIDI, "**Breast cancer diagnosis using an enhanced Extreme Learning Machine based-Neural Network",** 2018 International Conference on Signal, Image, Vision and their Applications (SIVA), Guelma, Algeria, 2018, pp. 1-4.

[5] M. Amrane, S. Oukid, I. Gagaoua and T. Ensarİ, "**Breast cancer classification using machine learning",** 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, 2018, pp. 1-4.

[6] 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.

[7] M. R. Al-Hadidi, A. Alarabeyyat and M. Alhanahnah, "**Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm",** 2016 9th International Conference on Developments in eSystems Engineering (DeSE), Liverpool, 2016, pp. 35-39.

[8] 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. 121-124.

[9] N. Jafarpisheh, N. Nafisi and M. Teshnehlab, **"Breast cancer relapse prognosis by classic and modern structures of machine learning algorithms",** 2018 6th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), Kerman, 2018, pp. 120-122.

[10] D. Bazazeh and R. Shubair,**"Comparative study of machine learning algorithms for breast cancer detection and diagnosis",** 2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, 2016, pp. 1-4.

[11] 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.

[12] Dua, D. and Graff, C. (2019). **UCI Machine Learning Repository** [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[13] Akinsola, J E T & Adeagbo, Moruf & Awoseyi, Ayomikun. (2020). "**Breast Cancer Predictive Analytics Using Supervised Machine Learning Techniques",** International Journal of Advanced Trends in Computer Science and Engineering. 8. 10.30534/ijatcse/2019/70862019.

[14] Arasi, Munya & Babu, Sangita. (2019). **"Survey of Machine Learning Techniques in Medical Imaging",** International Journal of Advanced Trends in Computer Science and Engineering. 8. 2017-2116. 10.30534/ijatcse/2019/39852019.

[15] Nallamala, Sri Hari & Mishra, Pragnyaban & Koneru, Suvarna. (2019). **"Qualitative Metrics on Breast Cancer Diagnosis with Neuro Fuzzy Inference Systems"**. International Journal of Advanced Trends in Computer Science and Engineering. 8. 259-264. 10.30534/ijatcse/2019/26822019.