IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 4, April 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104129 |

Survey on Breast Cancer Detection using Logistic Regression Algorithm

Prof. Ajit N.Gedam[1], Kajol B. Deshmane[2], Nishigandha N.Jadhav[3], Ritul M.Adhav[4], Akanksha N.Ghodake[5]

Dept. of Computer Science, AISSMS Polytechnic, Pune, MH, India[1,2,3,4,5]

ABSTRACT: Breast cancer is one of the common diseases specifically in women now days. It has become the second main reason of cancer death in females. Every year 4.5-5% new cancer cases are recorded and increasing the morbidity at worldwide. It has proved that early detection of any has proved that early detection of any cancer when followed up with appropriate diagnosis and treatment can increase the survival rate of the patients. Breast cancer is diagnosed by mammography. Mammograms are films generated by radiologist with a device. These mammograms are observed and diagnosed by the oncologist for further treatment. Since all general hospitals do not have the specialist and patients used to wait for their report. So, waiting for diagnosing a breast cancer may take time. This delay may be responsible for cancer spreading and reducing the survival rate of the patient. This does not mean to replace expert or physician by computer but it means that computer can assist the expert for better understanding the particular case and the results can be produced early. This survey paper presents a brief summary on breast cancer diagnosis using machine learning algorithms used to increase the efficiency and effectiveness of predicting cancer by comparing the different algorithms such as Logistic Regression, KNN, Naïve Bayes, Decision tree, SVM and random forest.So, from this survey we come to a conclusion that logistic regression algorithm is better and can be used to diagnose breast cancer by a computer to make the diagnosing efficient and effective. The correct diagnosis and accurate classification of breast cancer detection are the main objective of this survey paper.

KEYWORDS: Breast Cancer, Mammograms, machine learning algorithms, Logistic Regression.

I. INTRODUCTION

Over the years, a continuous evolution related to cancer research has been performed. Scientists used various methods, like early-stage screening, so that they could find different types of cancer before it could do any damage. With this research, they were able to develop new strategies to help predict early cancer treatment outcome. With the arrival of new technology in the medical field, huge amount of data related to cancer has been collected and is available for medical research. But physicians find the accurate prediction of the cancer outcome as the most.

In this project, I will be using different machine learning techniques to analyse the data given in the datasets. This analysis will help us predict whether the cancer is benign or malignant. Benign cancer is the cancer which doesn't spread whereas malignant cancer cells spread across the body making it very dangerous.

Machine learning algorithms will help with this analysis of the datasets. These techniques will be used to predict the outcome. The outcome can be either that the cancer is benign or malignant. Benign cancer is the cancer which doesn't spread whereas malignant cancer cells spread across the body making it very dangerous.

II. RELATED WORK

Sr. no.	Author Name	Paper Name	Advantage	Disadvantage
1.	Samer Hamed[1]; Abdelwadood Mesleh[2]; Abdullah Arabiyyat[3]	IJCSMC, Vol. 10, Issue. 11, November 2021, pg.4 – 11 Breast Cancer Detection Using Machine Learning Algorithms	1. It is simple and easy to implement. 2. It doesn't require as much training data. 3. It handles both continuous and discrete data. 4. It is highly scalable with the number of	1. Main imitation of Naive Bayes is the assumption of independent predictors. Naive Bayes implicitly assumes that all the attributes are mutually independent. 2. Its estimations

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104129 |

			predictors and data points. 5. It is fast and can be used to make real-time predictions.	can be wrong in some cases, so you shouldn't take its probability outputs very seriously.
2.	Shubham Sharma[1], Archit Aggarwal[2], Tanupriya Choudhury[3]	Breast Cancer Detection Using Machine Learning Algorithms	1. Simple to implement and intuitive to understand. 2. Can learn non-linear decision boundaries when used for classification and regression. 3. No Training Time for classification/regression: The KNN algorithm has no explicit training step and all the work happens during prediction.	**1.** Doesn't work well with a large dataset. **2.** Doesn't work well with a high number of dimensions. 3. Sensitive to outliers and missing values.
3.	M. Sumanth	International Journal of Multidisciplinary Engineering in Current Research Volume 6, Issue 12, December 2021 BREAST CANCER DETECTION WITH MACHINE LEARNING	1. A neural network can implement tasks that a linear program cannot. 2. When an item of the neural network declines, it can continue without some issues by its parallel features. 3. A neural network determines and does not require to be reprogrammed. 4. It can be executed in any application.	1. Hardware Dependence: Artificial Neural Networks require processors with parallel processing power, by their structure. For this reason, the realization of the equipment is dependent. 2. Unexplained functioning of the network: The most important problem of ANN. When ANN gives a probing solution, it does not give a clue as to why and how. This reduces trust in the network

1. **Samer Hamed[1]** focused on data mining techniques that are used to discover hidden patterns and their relationships. The author addresses the use of machine learning algorithm for predicting recurrence of breast cancer for the patients who were followed-up for two years. The scholar applied Iranian Centre for breast cancer (ICBC) to Support vector machine (SVM), Decision Tree (DT) and Artificial Neural Network (ANN). The Performances of these three techniques were compared and it was concluded that among the three SVM gives mores accuracy of 95.7%. Compared the result of two datasets Breast Cancer Coimbra (BCCD) and WBCD of breast cancer for DT, Random Forest (RF), SVM, Logistic Regression (LR) algorithms. The author computed the accuracy, measure metric of each algorithm in case of both data sets. It was concluded that SVM perform more accurately among all algorithms in both data sets with the accuracy of 97.7 and 76.3 respectively.

2. **Shubham Sharma[2],** kNN, Naïve Bayes and Random Forest have their advantage and disadvantage over each other in terms of performance, the type of problem they handle etc. As shown in kNN test time is O (1) without pre-processing of training set, in the case of Naïve Bayes: N is the number of training examples and d is the dimensionality of the features whereas for Random Forest [9]: N is the number of samples and K is the number of variables randomly drawn at each node. Naïve Bayes algorithm deals only with classification problems whereas both kNN and Random Forest can deal with classification as well as regression problems. In terms of accuracy both kNN and Random Forest can deliver high accuracy but Naïve Bayes algorithm need large number of records in order to yield

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104129 |

a better accuracy. Algorithms that simplify the function to a known form are called parametric machine learning algorithms, Naïve Bayes algorithm can be expressed as parametric as well as non-parametric model.

3. M. Sumanth[3], At DeSE 2016, the 9th International Conference on Developments in systems Engineering (DeSE), which took place in Liverpool in 2016, M. R. AlHadidi, A. Alarabeyyat, and M. Alhanahnah presented their findings. It was published at DeSE 2016, the 9th International Conference on Developments in Systems Engineering (DeSE), which was held in October. The study, "Breast Cancer Detection Using K-Nearest Neighbour Machine Learning Algorithm," was written by a team of researchers. As normal clinical practise in the medical field when it comes to diagnosing breast cancer, breast cancer detection pictures are used to aid in the identification of the disease.

Parameter	KNN	Naïve Bayes	Random Forest
Time Complexity (Training Phase)	O(1)	O(Nd)	$\Theta(MKN\log^2 N)$
Problem Type	Classification & Regression	Classification	Classification & Regression
Accuracy	Provides high accuracy	For high accuracy it needs very large number of records	Provides high accuracy
Model Parameter	Non Parametric	Parametric/Non Parametric	Non Parametric

Table 1. Comparison among KNN, Naïve Bayes and RandomForest

- ***Limitations of KNN:***

KNN is a very powerful algorithm. It is also called “lazy learner”. However, it has the following set of limitations:

1. **Doesn’t work well with a large dataset:**
Since KNN is a distance-based algorithm, the cost of calculating distance between a new point and each existing point is very high which in turn degrades the performance of the algorithm.

2. Doesn’t work well with a high number of dimensions:
Again, the same reason as above. In higher dimensional space, the cost tocalculate distance becomes expensive and hence impacts the performance.

4. **Sensitive to outliers and missing values:**
KNN is sensitive to outliers and missing values and hence we first need to impute the missing values and get rid of the outliers before applying the KNN algorithm.

- **Limitations of Naïve Bayes:**

1. Main imitation of Naive Bayes is the assumption of independent predictors. Naive Bayes implicitly assumes that all the attributes are mutually independent. In real life, it is almost impossible that we get a set of predictors which are completely independent.
2. If categorical variable has a category in test data set, which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as Zero Frequency.
3. Its estimations can be wrong in some cases, so you shouldn’t take its probability outputs very seriously.

- **Limitation of ANN:**

1. Hardware Dependence: Artificial Neural Networks require processors with parallel processing power, by their structure. For this reason, the realization of the equipment is dependent.
2. Unexplained functioning of the network: The most important problem of ANN. When ANN gives a probing solution, it does not give a clue as to why and how. This reduces trust in the network.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104129 |

3. Assurance of proper network structure: There is no specific rule for determining the structure of artificial neural networks. The appropriate network structure is achieved through experience and trial and error.
4. The difficulty of showing the problem to the network: ANNs can work with numerical information. Problems have to be translated into numerical values before being introduced to ANN. The display mechanism to be determined will directly influence the performance of the network. This is dependent on the user's ability.

- **Limitations of Decision Tree algorithm:**

1. A small change in the data can cause a large change in the structure of the decision tree causing instability.
2. For a Decision tree sometimes calculation can go far more complex compared to other algorithms.
3. Decision tree often involves higher time to train the model.
4. Decision tree training is relatively expensive as the complexity and time has taken are more.

- **Limitations of SVM:**

1. SVM algorithm is not suitable for large data sets.

2. SVM does not perform very well when the data set has more noise i.e., target classes are overlapping.

3. In cases where the number of features point exceeds the number of trainingdata samples, the SVM will underperform.

4. As the support vector classifier works by putting data points, above and below the classifying hyper plane there is no probabilistic explanation for the classification.

IV. METHODOLOGY

- **Logistic model:**

Logistic regression was introduced by statistician DR Cox in 1958 and so predates the field of machine learning. It is a supervised machine learning technique, employed in classification jobs (for predictions based on training data). Logistic Regression uses an equation likeLinear Regression, but the outcome of logistic regression is a categorical variable whereas it is a value for other regression models. Binary outcomes can be predicted from the independent variables.
The general workflow is:
(1) get a dataset
(2) train a classifier
(3) make a prediction using such classifier

Main working of model:

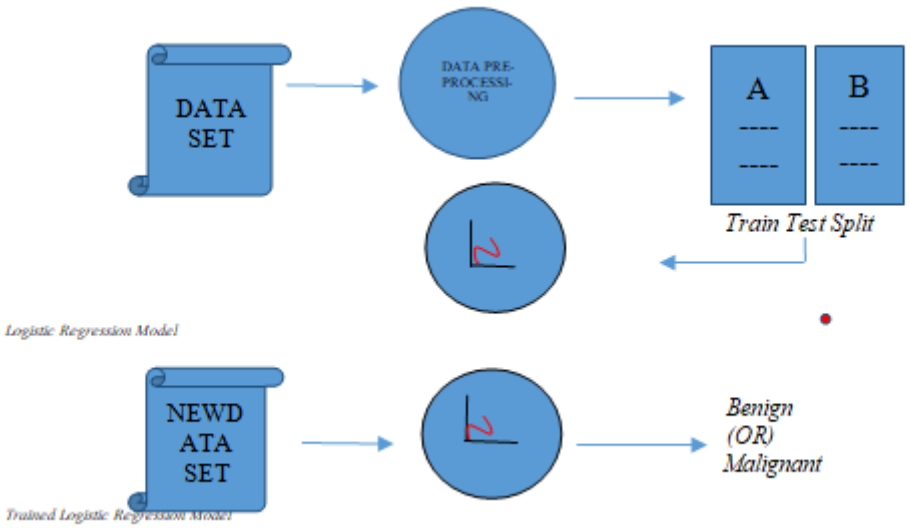


Fig: 1.1Work flow diagram for breast cancer detection

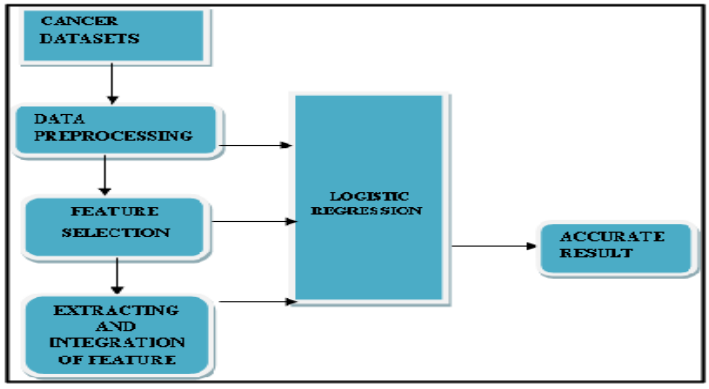
| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

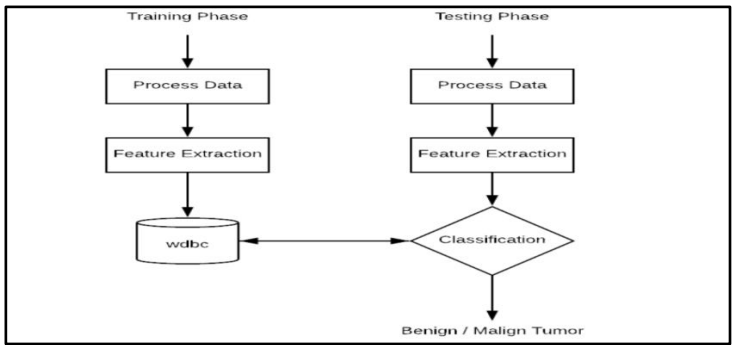
| DOI:10.15680/IJIRSET.2022.1104129 |

So according to the about work flow diagram first one data set will be used, after that the data set will move towards data pre-processing (Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format) and in this process the data set will be compatible for python language, and the data set will be trained.

After this data is pre-processed, we can use the data set in the logistic regression algorithm which is the machine learning algorithm. Then we can choose any new random value from the data set which will be tested in the trained logistic regression algorithm where it will help us to detect whether the patient has Benign or Malignant type of cancer.



1.2 Flow Chart of Logistic Regression



1.3 Architecture diagram for breast cancer detection.

IV. CONCLUSION

The most frequently occurring type of across cancer is breast cancer. There is a chance of twelve percent for a women picked randomly to be diagnosed with the disease. Thus, early detection of breast cancer can save a lot of valuable life. The proposed work flow in our project is the study of machine learning algorithms, for the detection of breast cancer. It has been observed that logistic regression algorithm had an accuracy of more than 94%, to determine benign cancer or malignant cancer. Thus, supervised machine learning techniques will be very supportive in early diagnosis and prognosis of a cancer type in cancer research.

REFERENCES

1. Sameer Hamed1; Abdelwadood Mesleh2; Abdullah Arabiyyat3, Vol. 10, Issue. 11, November 2021, pg.4 – 11 Breast Cancer Detection Using Machine Learning Algorithms
2. Shubham Sharma[1], Archit Aggarwal[2], Tanupriya Choudhury[3], Breast Cancer Detection Using Machine Learning Algorithms
3. International Journal of Multidisciplinary Engineering in Current Research Volume 6, Issue 12, December 2021 BREAST CANCER DETECTION WITH MACHINE LEARNING

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104129 |

4. Breast Cancer Detection and Prediction using Machine Learning Anoy Chowdhury University of Engineering & Management, Kolkata.
5. 2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3) Mody University of Science and Technology, Lakshmangarh, Feb 21-22, 2020 978-1-7281-1420-0/20/$31.00 ©2020 IEEE Diagnosis of Breast Cancer using Machine Learning Algorithms-A Review
6. M. R. Basunia, I. A. Pervin, M. A. Mahmud, S. Saha, and M. Arifuzzaman, "On Predicting and Analyzing Breast Cancer using Data Mining Approach," in 2020 IEEE Region 10 Symposium (TENSYMP), 2020, pp. 1257-1260.
7. P. Mekha and N. Teeyasuksaet, "Deep Learning Algorithms for Predicting Breast Cancer Based on Tumour Cells," in 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON), 2019, pp. 343-346.
8. S. Kabir and M. Ahad, "ANN Classification of Female Breast Tumour Type Prediction Using EIM Parameters," in 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), 2020, pp. 890-893.
9. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," CA Cancer J Clin, vol. 71, pp. 209-249, May 2021.
10. M. Masud, A. E. EldinRashed, and M. S. Hossain, "Convolutional neural network-based models for diagnosis of breast cancer," Neural ComputAppl, pp. 1-12, Oct 9 2020.
11. H. Abdel-Razeq, A. Mansour, and D. Jaddan, "Breast Cancer Care in Jordan," JCO global oncology, vol. 6, pp. 260-268, 2020.
12. "Breast Cancer Screening and Diagnosis," Annals of Internal Medicine, vol. 172, pp. 839-840, 2020.
13. G. G. Hungilo, G. Emmanuel, and A. W. R. Emanuel, "Performance Evaluation of Ensembles Algorithms in Prediction of Breast Cancer," in 2019 International Biomedical Instrumentation and Technology Conference (IBITeC), 2019, pp. 74-79.
14. V. A. Telsang and K. Hegde, "Breast Cancer Prediction Analysis using Machine Learning Algorithms," in 2020 International Conference on Communication, Computing and Industry 4.0 (C2I4), 2020, pp. 1-5.
15. K. Sharma, B. Muktha, A. Rani, and C. Thirumalai, "Prediction of benign and malignant tumor," in 2017 International Conference on Trends in Electronics and Informatics (ICEI), 2017, pp. 1057-1060.
16. F. S. Ahadi, M. R. Desai, C. Lei, Y. Li, and R. Jia, "Feature-Based classification and diagnosis of breast cancer using fuzzy inference system," in 2017 IEEE International Conference on Information and Automation (ICIA), 2017, pp. 517-522.
17. P. Sivakumar, T. U. Lakshmi, N. S. Reddy, R. Pavani, and V. Chaitanya, "Breast Cancer Prediction System: A novel approach to predict the accuracy using Majority-Voting Based Hybrid Classifier (MBHC)," in 2020 IEEE India Council International Subsections Conference (INDISCON), 2020, pp. 57-62.
18. N. A. Mashudi, S. A. Rossli, N. Ahmad, and N. M. Noor, "Comparison on Some Machine Learning Techniques in Breast Cancer Classification," in 2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), 2021, pp. 499-504.
19. S. Ara, A. Das, and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," in 2021 International Conference on Artificial Intelligence (ICAI), 2021, pp. 97-101.
20. M. R. Basunia, I. A. Pervin, M. A. Mahmud, S. Saha, and M. Arifuzzaman, "On Predicting and Analyzing Breast Cancer using Data Mining Approach," in 2020 IEEE Region 10 Symposium (TENSYMP), 2020, pp. 1257-1260.
21. P. Mekha and N. Teeyasuksaet, "Deep Learning Algorithms for Predicting Breast Cancer Based on Tumor Cells," in 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON), 2019, pp. 343-346.
22. S. Kabir and M. Ahad, "ANN Classification of Female Breast Tumor Type Prediction Using EIM Parameters," in 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), 2020, pp. 890-893.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details