


Review Article

A Review on Breast Cancer Prediction Using Machine Learning

*Swasti Goyal ¹, Tanya Sharma², Anuj Kumar³

¹Research Scholar, Dept. of Computer Science & Engineering, Shobhit University, Gangoh, India
swastigoyal@gmail.com <https://orcid.org/0000-0002-6982-162X>

^{2,3}Asst. Professor, Dept. of Computer Science & Engineering, Shobhit University, Gangoh, India

*Corresponding Author – swastigoyal@gmail.com

DOI - <https://doi.org/10.55083/irjeas.2022.v10i04001>

© 2022 Swasti et al.

This is an article under the CC-BY license. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. <https://creativecommons.org/licenses/by/4.0/>

Received: 06 September 2022; Received in revised form: 04 October 2022; Accepted: 13 October 2022

Abstract: One of the most in demand research topic in today's technology world is medical area and cancer is one of them. The second main cause of death in the world is cancer. In 2015 about 8.8 million people have died due to cancer. For the early detection of breast cancer several types of research have been done to start the treatment and increase the survivability. Breast cancer affects the women mentally as well as emotionally. The goal of this challenge is to provide a framework that uses a cancer medical dataset as input and then analyses the dataset to produce findings that help medical experts better understand the state of the disease. The majority of studies concentrate on mammography results. However, incorrect detection in mammography pictures can occasionally result, endangering the patient's health. Cancer that forms in the cells of breast is said to be breast cancer. The cancer should be cured if it is diagnosed in early stage. So, with the help of machine learning algorithm the cancer should be diagnosed early. Most of the women lives are affected by the breast cancer in all over the world.

There are two types of tumors that can be found in breast cancer i.e. malignant or benign. If a person is having a cancer disease, then it will be categorized as malignant otherwise it is known as benign. In 2020, 685000 deaths and 2.3 million women diagnosed with breast cancer globally, somewhere in world in every 14 seconds, a woman is diagnosed with breast cancer [1]. Patient life from the breast cancer can be saved only if it is found in early stage; if it is diagnosed later then the chances of survival are less. If the cancer is diagnosed early then the patient will get a better treatment. This study will concentrate on a few machine learning methods for identifying if a breast cancer is malignant or benign. The Wisconsin Breast Cancer Dataset, which was acquired via Kaggle, was used in this study. Our goal is to evaluate how accurately various machine learning algorithms can detect breast cancer. These include Random Forest Classifier, Decision Tree Classifier, Support Vector Machine, and K-Nearest Neighbors. All the experiments are conducted on a Jupiter platform. After analyzing the accuracy of each algorithm the most suitable one is Support Vector Machine that gives the better accuracy among all i.e., 98%.

Keywords: Breast Cancer, KNN, Machine Learning, SVM, Tumor

1. INTRODUCTION

Cancer, particularly breast cancer in women, is one of the most hazardous diseases in the world. Breast cancer causes the deaths of numerous women each year.

Breast cancer can be found physically, but it will take a while. Therefore, it will make it difficult for doctors to diagnose diseases. Globally, breast cancer is the leading cause of mortality for women. By 2030, such chronic diseases are expected to claim the lives of over 23.6 million

people, according to the World Health Organization (WHO) [1]. Breast cancer is the most frequent cancer in women and has the highest cancer death rate in Malaysia, where it is confirmed that Malaysian women with breast cancer present with the disease at a later stage than women in other nations [18]. In Malaysia, 5% of women are thought to be at risk for breast cancer, compared to 12.5% in Europe and the US [19]. If symptoms emerge, breast cancer can be quickly and easily detected. On the other hand, many women with breast cancer do not exhibit any symptoms. As a result, routine breast screening is crucial for early detection [20]. Exceedingly few people are able to receive therapy, and for those that do, it is usually very expensive and difficult. Furthermore, taking a lengthy time and delaying decisions may be a factor in fatalities. Breast cancer is a chronic condition that is difficult to treat. Therefore, the automatic identification of cancer using multiple diagnostic procedures is critically important. Because of this, the majority of patients are unable to pay for the cost of treating their cancer. Breast cancer is fairly detectable if symptoms start to occur. However, many women who have breast cancer don't exhibit any symptoms. Breast cancer is the most prevalent type of cancer in women, both in terms of frequency and mortality. Because of researchers' efforts and a variety of early detection techniques, the temporality rate has been declining over the past few decades. According to different studies, machine learning provides a higher level of accuracy in matters relating to the medical area. Over the past few years, breast cancer diagnosis has seen a significant increase in the usage of machine learning approaches. The accuracy of a patient's diagnosis relied on the doctor's prior expertise. Although the diagnosis was established over many years of observation and investigation of various patient symptoms, accuracy could not be guaranteed, as computer technology advanced [2]. Breast cancer typically affects women in their 40s and older. Breast cancer should develop when the lobules, the milk-producing cells in the gland, are aberrant and divide quickly. In Malaysia, breast cancer is the most common type of cancer that should affect women [3]. A hard mass or lump in the breast tissue is a sign of breast cancer. Only 78% of women who have mammograms get reliable diagnoses. The American Cancer Society recommends that all women start annual mammography at age 40, contrary to the U.S. Preventive Services Task Force (USPSTF), which suggests getting a baseline mammogram at age 40 and starting yearly testing at age 50 [4]. The various machine learning algorithms were employed for the diagnosis of both

malignant and benign breast cancer. Recovery from breast cancer is a lengthy process. According to (WHO) [1], there would be a death rate of about 23.6 million people by 2030]. The 98% survival percentage for breast cancer can be increased with early detection [5]. The most prevalent malignancy among women worldwide is breast cancer. It is brought on by some breast cells growing abnormally. There are various types of breast cancer, and these cancerous cells can travel throughout the body and harm the entire body.

Breast cancer is a condition that affects people on a regular basis in women between the ages of 40 and 50. Three diseases include lymphoma, melanoma, and malignant thyroid development. You can alter other risk factors, like smoking. There is a need of Computer Aided Diagnosis system using machine learning approach for the diagnosis of breast cancer due to its extremity. We are using various machines learning algorithm to check the accuracy of different algorithms. From Decision tree classifier 0.92 accuracy have been achieved, 0.95 accuracy from Random Forest classifier, 0.95 from K-Nearest Neighbors, and 0.98 from Support Vector Machine. Therefore, we are getting the higher accuracy from Support Vector Machine after comparing all the algorithms.

2. LITERATURE REVIEW

Naqa et al. employ 1120 MCs from 76 clinical mammograms. The primary goal of this support vector machine application is to find clusters of microcalcifications in digital mammograms. It is founded on the idea of minimizing risk [6]. For the diagnosis of breast cancer, Maglogiannis et al. employ the Wisconsin Prognostic Breast Cancer Dataset using the SVM algorithm [7]. In order to predict cancer, Thongkam et al. used Adaboost and random forest. The author believes that random forest provides the best accuracy, and in the future, ARBF algorithm was utilised [8]. While computing future ROC values using the WDBC dataset, Mert et al. employ SVM, which provides a greater accuracy of 94.1%; however, Srudy states that accuracy is decreased when using ICA. [9]. Hussain et al. uses 10-fold cross validation for accuracy, sensitivity, and specificity. The main purpose of using different SVM kernel to compare the accuracy and use the evaluation on 5X2 cross validation SVM Mahalanobis gives best accuracy and in future different kernels were combined to give better accuracy [10]. Nematzadeh et al. uses WBCD and WDBC dataset where they compare DT, MLP, SMO algorithm where SMO gives the higher accuracy [11, 12]. On the Wisconsin Breast Cancer

Dataset, Bazazeh et al. compare the precision of three machine learning techniques. According to the research Ann has the higher accuracy [13]. Islam et al. uses SVM and KNN to compare accuracy of algorithm where SVM gives the best accuracy and in future larger dataset is used [14]. Sengar et al. uses two algorithms i.e., Decision tree classifier and Logistic Regression. Decision tree gives the best accuracy, they trained their machine using multivariate regression, they are using 570 rows and 32 columns [16]. Ara et al. collected data from Fine Needle Aspiration (FNA) and using SVM, LR, KNN, DT, NB, RF

and compare these algorithms whereas, support vector machine gives the higher accuracy [17]. Velliangiri et al. uses dataset from UCI repository and compares different machine learning algorithm such as LR, SVM, NN, RF, KNN, and LDA where LDA gives the higher accuracy i.e., 97% in future DICOM images were used to predict breast cancer. Divya et al. used dataset from UCI repository, DICOM images were used in future to predict cancer.

Summary of Literature Review

<u>Reference</u>	<u>Methodology</u>	<u>Dataset</u>	<u>Results</u>	<u>Limitation</u>	<u>Future scope</u>
[6] 2002	SVM	The University of Chicago and Department of Radiology	SVM 94%	More training time should be need	The effectiveness of the trained classifier can be further enhanced by the SEL scheme.
[7] 2006	(SVM)	(WDBC) and the (WPBC) datasets	SVM (96.91%)	Probably,svm doesn't work well with larger dataset as requires higher training time.	An efficient voting classification system should be deployed to find out best perform algorithm.
[8] 2008	AdaBoost , ABRF, C4.5, and RF	Database- WEKA Dataset- Srinagarind hospital in thailand	AdaBoost 88.90% RF 94.30% ABRF 94.40% C4.5 85.20%	Can make the prediction process slow	Utilizing the ABRF technique in larger data sets for investigation
[9] 2011	(SVM)	(WDBC) data set	(SVM) 94.41%	ICA will decrease-s the accuracy and sensitivity Values	In addition, the receiver operating characteristic (ROC) curve and its criterion values have computed
[10] 2011	SVM	WBCD from the University of California's Machine Learning Repository	SVM RBF 96.74%, SVM polynomial 96.79%, SVM Mahala Nobis 97.06%, and SVM sigmoid 96.13%	The performance is not considerably impacted by the feature subset selection.	research into combining various kernels for better detection outcomes

[11] 2012	DT (J48), (MLP), (NB), (SMO) base on knn.	Database- WEKA Dataset- WBCD, WDBC	NB 95.9943 % MLP 95.279 % J48 95.1359 % SMO 96.9957 % IBK 94.563	Combination of different algorithm is difficult to analyze	Use of images to predict breast cancer
[12] 2015	DT, NB, NN, SVM	(WBC), (WDBC) and (WPBC)	Svm-linear 97.63 Svm-mlp 96.13 Svm-rbf 95.91 NN 98.09 Decision Tree 93.35 Naïve Bayes 97.14	May be out of memory error occurred	The Genetic Algorithm and Particle Swarm Optimization methods can be used to arrive at the value of k.
[13] 2016	(SVM), (RF) and (BN)	Database- WEKA Dataset- WBCD	Recall values SVM97.0% RF96.6% BN97.1% Precision values SVM97.0% RF96.6% BN97.2% Area under ROC values SVM96.6% RF99.9% BN99.1%	It can handle only small datasets	May be images were used to predict breast cancer
[14] 2017	SVM, KNN	Database- WEKA Dataset- WBCD set from UCI machine learning repository dataset	SVM 98.57% KNN 97.14%	Doesn't work well when dataset has more noise	Images were added in future to predict the cancer.
[15] 2019	NB, J48, RBF, SVM	WDBC from an online data mining repository of the University of California(UCI)	NB 97% J48 93.41% RBF 96.77% SVM 97.07%	Doesn't work with larger dataset	Compare the data with larger dataset
[16] 2020	(LR), and	WDBC	LR 0.94%	Sometime calculation become	Compare the dataset with more

	(DT)		DT 0.95%	more complex	algorithms to predicts better accuracy
[17] 2021	SVM, LR, MLR, KNN, DT, RF, NB	Wisconsin Breast Cancer Dataset from UCI repository	LR 94.4% KNN 95.8% DT 95.1% NB 92.3% RF 96.5% SVM 96.5%	It is very sensitive to unrelated features	Add more function and work on larger dataset
[18] 2022	L.R., SVM, N.N., R.F., KNN, and LDA	dataset from the UCI repository	LR 88.75 SVM 94 NN 65 RF93 KNN 91.21 LDA 97	Check the accuracy with other techniques for better output	DICOM images is used to predict breast cancer

3. FINDING AND GAPS

- According to (Naqa et al., 2002) only svm is used more algorithms are used in future to predict better outcome [6].
- According to our study (prateek et al., 2020).al (LR), and (DT) were used and in this calculation becomes more complex [16].
- According to our study (Thongkam et al., 2008) are using AdaBoost this makes the prediction process slow but this should be able to extracting methods [8].
- According to literature review (Bazazeh et al., 2016) uses (SVM), (RF) and (BN) were used to predict breast cancer but in future images were used to predict [13].
- According to our study (Yang et al., 2002) predicts that SVM gave the higher accuracy in most of the research paper as after comparing the different algorithms [6] [7] [13] [14] [15] .
- Author to our literature review (Yen et al., 2022) uses L.R., SVM, N.N., R.F., KNN, and LDA in LDA gives the higher accuracy and DICOM images were used in future to predict breast cancer [18].
- According to our study WHO WBCD is used in various algorithms so, in future large dataset were used for predicting breast cancer [17].

4. METHODS USED

In this setup we are using machine learning techniques for the intelligent breast cancer detection. This setup focuses on four steps to make a decision. How to extract data and combined data from different health system that what's the phase will do, and phase two includes how to store large amount of medical data. To train the data of cancer disease dataset machine learning based classifier approach should be used in phase three. Therefore, phase four predict the output for the breast cancer detection client.

We are using the following machine learning classifier-:

- **SVM**

The main objective of the SVM approach is to locate the hyperplane in an N-dimensional space that clearly classifies the data points. The attributes that we entered determine the hyperplane's dimensions. Suppose, if we are entering only two features then the hyperplane is only a line, therefore if we entered three features then the hyperplane will be 2-D. Implementing SVM is more different as compared to other machine learning algorithms. It can handle multiple continuous values.

- **Decision Tree Classifier**

Decision node and leaf node are nodes which are present in the decision tree. To make any decision, decision node is used and having multiple branches, the decision node that do not contain any further branch then there output will be leaf node. It starts with a root node and expands with further branches and builds a tree like structure that's why it is called decision tree.

- **Random Forest**

That contains a various number of decision tree on a various subset of given dataset is said to be Random forest classifier. It should improve the accuracy of dataset. Each tree's output is predicted by the random forest, which may then anticipate the overall outcome based on the majority of predictions.

- **K-Nearest Neighbors**

K-Nearest Neighbors forecast datapoint values and further award fresh datapoints a score based on how closely they resemble points in the training dataset.

• Data Collection

In this study, the Wisconsin Breast Cancer Dataset, which was obtained from Kaggle, was used. Breast cancer dataset has 569 rows and 33 columns, but

only 357 benign and 212 malignant cells are present.

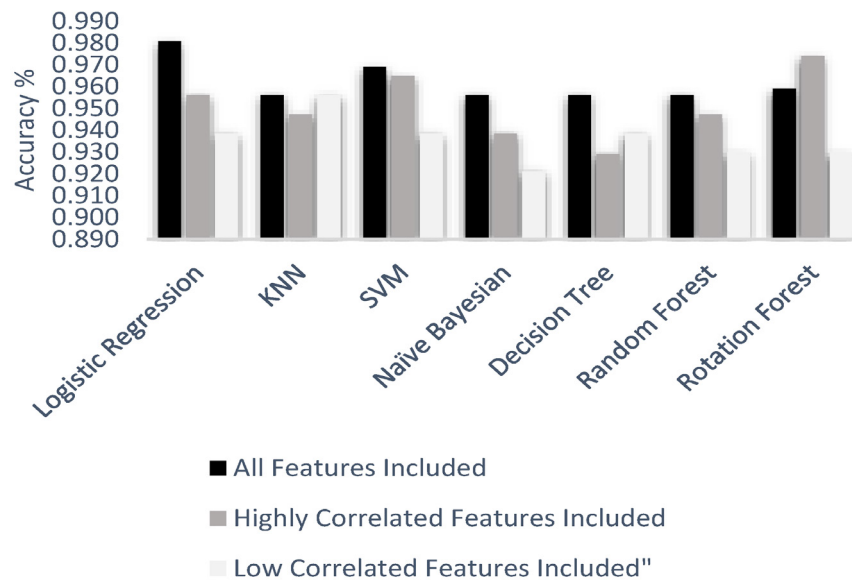


Figure 1- Comparative result for detection of Breast Cancer

5. CONCLUSION & FUTURE SCOPE

Most of the women deaths occur in all over the world because of breast cancer. Without the aid of machine learning techniques, early detection is impossible. So, in order to assess whether the tumor is malignant or benign, several machine learning techniques are employed. We are trying to finding the best result through the Support Vector Machine Classifier. But some other techniques were used to get the better output. DT gives the accuracy of 0.92% whereas RF and KNN gives the accuracy of 0.95% and we are obtaining the higher accuracy from SVM i.e., 0.98%. From these results, we can understand that SVM performs better and give the better prediction.

In the future, we may be able to predict breast cancer using deep learning techniques or photos, improving both diagnosis and prediction. The goal is to maximize accuracy while lowering the mistake rate.

REFERENCES

- [1] WHO (World Health Organization) Online: <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>.
- [2] Meesad, P, Yen, G.G. Combined numerical and linguistic knowledge representation and its application to medical diagnosis. IEEE Trans. Syst. Man Cybern. 2003, 33,206-222.
- [3] About Breast Cancer[Online: <https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470>].
- [4] UCHealth 2015," How accurate are mammograms?" [Online: <https://www.uchealth.org/today/how-accurate-are-mammograms/>].
- [5] S. A. Korkmaz and M. Poyraz, "A New Method Based for Diagnosis of Breast Cancer Cells from Microscopic Images: DWEE—JHT," Journal of Medical Systems, vol. 38, no. 9. Springer Science and Business Media LLC, Jul. 15, 2014. doi: 10.1007/s10916-014-0092-3.
- [6] A Support Vector Machine Approach for Detection of Microcalcifications Issam El-Naqa, Student Member, IEEE, Yongyi Yang, Member, IEEE, Miles N. Wernick, Senior Member, IEEE, Nikolai P. Galatsanos, Senior Member, IEEE, and Robert M. Nishikawa 2002.
- [7] An Intelligent System for Automated Breast Cancer Diagnosis and Prognosis using SVM based Classifiers Ilias Maglogiannis, Elias Zafiroopoulos, and Ioannis Anagnostopoulos 2006.

- [8] AdaBoost Algorithm with Random Forests for Predicting Breast Cancer Survivability Jaree Thongkam, Guandong Xu and Yanchun Zhang 2008.
- [9] Breast Cancer Classification by Using Support Vector Machines with Reduced Dimension Ahmet Mertl , Niyazi Kilic, Aydn Akan 2011.
- [10] A Comparison of SVM Kernel Functions for Breast Cancer Detection Muhammad Hussain, Summrina Kanwal Wajid, Ali Elzaart, Mohammed Berbar 2012
- [11] Comparative Studies on Breast Cancer Classifications with K-Fold Cross Validations Using Machine Learning Techniques Zahra Nematzadeh, Roliana Ibrahim, Ali Selamat 2015
- [12] Md. M. Islam, Md. R. Haque, H. Iqbal, Md. M. Hasan, M. Hasan, and M. N. Kabir, "Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques," SN Computer Science, vol. 1, no. 5. Springer Science and Business Media LLC, Sep. 2020. doi: 10.1007/s42979-020-00305-w.
- [13] Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis Dana Bazazeh and Raed Shubair 2016
- [14] Prediction of breast cancer using support vector machine and K-Nearest Neighbors MM Islam, H Iqbal, MR Haque 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 226-229, 2017
- [15] Breast cancer via machine learning Mamatha Sai Yarabarla, Lakshmi Kavya Ravi, A Sivasangari 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 121-124, 2019
- [16] Comparative study of machine learning algorithms for breast cancer prediction Prateek P Sengar, Mihir J Gaikwad, Ashlesha S Nagdive 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 796-801, 2020
- [17] M. Kalaiyarasi, R. Dhanasekar, S. Sakthiya Ram, and P. Vaishnavi, "Classification of Benign or Malignant Tumor Using Machine Learning," IOP Conference Series: Materials Science and Engineering, vol. 995, no. 1. IOP Publishing, p. 012028, Nov. 01, 2020. doi: 10.1088/1757-899x/995/1/012028.
- [18] P. Divya, D. Palanivel Rajan, R. Suguna, S. Velliangiri 2022 International Conference on Computer Communication and Informatics (ICCCI - 2022), Jan. 25-27, 2022, Coimbatore, INDIA 978-1-6654-8035-2/22/\$31.00 ©2022 IEEE Machine Learning Techniques for Prediction and Analysis of Benign and Malignant in Breast Cancer
- [19] T.M Khan, S.A Jacob, Brief review of complementary and alternative medicine use among Malaysian women among breast cancer, Journal of pharmacy practice and research 2017.
- [20] R. Kirubakaran, T. chee jia, and N. Mahamad Aris, "Awareness of Breast Cancer among Surgical Patients in a Tertiary Hospital in Malaysia," APJCP, vol. 18, no. 1, Jan. 2017, doi: 10.22034/APJCP.2017.18.1.115.

Conflict of Interest Statement: The authors declare that there is no conflict of interest regarding the publication of this paper.

Copyright © 2022 Swasti Goyal, Tanya Sharma, Anuj Kumar. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

This is an open access article under the CC-BY license.
Know more on licensing on

<https://creativecommons.org/licenses/by/4.0/>

